Steep Turn On/Off "Green" Tunnel Transistors

Pratik Ashvin Patel

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-164
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-164.html

December 17, 2010

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **17 DEC 2010**	2. REPORT TYPE	3. DATES COVERED **00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE **Steep Turn On/Off ’Green’ Tunnel Transistors**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Department of Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Scaling of supply voltage Vdd has significantly slowed down since the 130 nm node. As a result, integrated circuit (IC) power consumption has been on the rapid rise. This presents a serious thermal management challenge and potential limiter of integration density as well as a rapidly growing portion of the world electricity demand. The problem lies in the 60 mV/dec swing limitation of any device involving charge flow over energy barrier (i.e., current state of art CMOS). This requires at least 60 mV to decrease the transistor current by 10X. The future low power or ?green? energy efficient scenario would benefit from a device that is friendlier to Vdd scaling. A transistor where carriers tunnel through rather than flow over a barrier is not subject to this limitation. However, achieving sub 60 mV/dec at current ranges of interest and over many decades is not trivial when relying solely on transmission probability modulation (i.e. increase/decrease of tunnel barrier width). Instead, if the absence/presence of tunneling state overlap is exploited a sharp ?off? to ?on? transition is achievable. By engineering the transistor device structure such that this overlap (i.e., onset of tunneling) occurs in a region of high electric field results in steep sub 60 mV/dec response over many decades of current. One novel design utilizes heavily doped, ultra shallow N+/P+ junctions to achieve this "sudden tunneling overlap" effect. Another design involves use of ultra thin body silicon-on-insulator (5 nm) to achieve a similar effect. Simulation results show sub 500 mV Vdd is possible if suitable low-Eg material is introduced. Both designs have been fabricated in silicon and their measurement results are presented.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**106**	

Steep Turn On/Off “Green” Tunnel Transistors

By

Pratik Ashvin Patel

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Chenming Hu
Professor Tsu-Jae King Liu
Professor Ronald Gronsky

Fall 2010

Abstract

Steep Turn On/Off “Green” Tunnel Transistors

by

Pratik Ashvin Patel

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Chenming Hu, Chair

Scaling of supply voltage V_{dd} has significantly slowed down since the 130 nm node. As a result, integrated circuit (IC) power consumption has been on the rapid rise. This presents a serious thermal management challenge and potential limiter of integration density as well as a rapidly growing portion of the world electricity demand. The problem lies in the 60 mV/dec swing limitation of any device involving charge flow over energy barrier (i.e., current state of art CMOS). This requires at least 60 mV to decrease the transistor current by 10X. The future low power or “green” energy efficient scenario would benefit from a device that is friendlier to V_{dd} scaling. A transistor where carriers tunnel through rather than flow over a barrier is not subject to this limitation. However, achieving sub 60 mV/dec at current ranges of interest and over many decades is not trivial when relying solely on transmission probability modulation (i.e., increase/decrease of tunnel barrier width). Instead, if the absence/presence of tunneling state overlap is exploited a sharp “off” to “on” transition is achievable. By engineering the transistor device structure such that this overlap (i.e., onset of tunneling) occurs in a region of high electric field results in steep sub 60 mV/dec response over many decades of current. One novel design utilizes heavily doped, ultra shallow N+/P+ junctions to achieve this "sudden tunneling overlap" effect. Another design involves use of ultra thin body silicon-on-insulator (5 nm) to achieve a similar effect. Simulation results show sub 500 mV V_{dd} is possible if suitable low-E_g material is introduced. Both designs have been fabricated in silicon and their measurement results are presented.

Table of Contents

Acknowledgements

Firstly, I would like to express my most sincere thanks to my research advisor Prof. Chenming Hu. His biggest advice to me was that good research involves asking the right questions. I have become a better engineer, researcher, and scholar as result of his simple advice. His constant energy, enthusiasm, and interest in my work have been unwavering for the past 6 years. I would also like to thank Prof. Tsu-Jae King Liu for being a member of my committee and for all her advice and feedback during our device group meetings. Prof. King has been a wonderful source of help with regards to fabrication challenges in Microlab for all device group students. I would like to thank Prof. Ronald Gronsky for being a member of my dissertation committee and helping to review my thesis.

The experimental results presented in this thesis would not be possible without all of Microlab staff. They are charged with the impossible task of keeping the tools in Microlab and the new Nanolab in order. Generations of device group students have benefited from their expertise and diligent hard work. I would like to thank Dr. Prashant Majhi for the opportunity to work with Sematech, our collaboration partner for this project. I would like to thank Dr. Greg Smith, who was a great friend and mentor at Sematech.

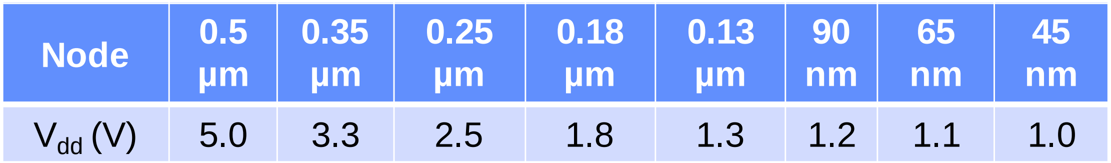
I would like to acknowledge MARCO MSD and DARPA STEEP for funding and supporting our research.

I would also like to thank my family for teaching me how to learn and the joy of learning from a very young age. They have supported me and given me the opportunity to continue learning for the past 29 years. I would not be where I am or who I am today without their support and encouragement.

Lastly, I would like to thank device group, one of the smartest and friendliest collaborative research groups in the world. I consider it a privilege to be a device group alumnus. I cannot imagine my time in graduate school without my fellow device group colleagues, who I have learned so much from. I want to thank, in particular, Cheuk Chi Lo and Anupama Bowonder for being my closest friends. We entered the same year and have been through the worst of times and best of times in graduate school together. Cheuk has always been willing help me and more than happy to answer my questions even when I bother him too much. I will miss playing tennis on Fridays. Finally, I want to thank Anupama for all her help, support, and love. I am incredibly lucky to have met my wonderful fiancée in graduate school.

Chapter 1: The Need for "Greener" Transistors

1.1 Losing an Effective Handle on Power Consumption



Node	0.5 µm	0.35 µm	0.25 µm	0.18 µm	0.13 µm	90 nm	65 nm	45 nm
V_{dd} (V)	5.0	3.3	2.5	1.8	1.3	1.2	1.1	1.0

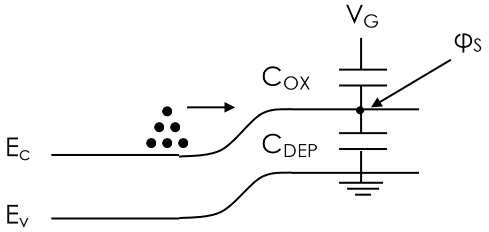
* ITRS Roadmap

Figure 1.1: Historical scaling trends of supply voltage vs. technology node. [1.1]

With each semiconductor technology node, transistors are made smaller and faster. The number of transistor per chip or integration density increases. The performance or chip clock frequency also increases as a result of faster devices and smaller capacitance. This suggests that the chip power density, which is defined as the dynamic power consumption per area, would increase with each technology node. However, as seen in Figure 1.1 the supply voltage V_{dd} has been decreasing with each node. V_{dd} reduction is the most effective handle on power consumption of integrated circuits. The dynamic power consumption is proportional to the V_{dd}^2.

$$P_{dynamic} \propto CV_{dd}^2 \tag{1.1}$$

Historically, the V_{dd} has been reduced in relation to the node feature size. For example, starting at 0.5 µm node the V_{dd} was 5.0 V, 3.3 V for 0.35 µm, 2.5 V for 0.25 µm up until 1.3 V for the 0.13 µm. However, starting at the 90 nm node supply voltage scaling has slowed down. Currently the 45 nm node uses V_{dd} of 1.0 V, instead of a projected 0.45 V, implying that current integrated circuits are consuming 4X more power than expected. In actuality, starting at 65 nm node significant circuit architectural changes have been made to curtail power density and permit scaling to continue, such as multi-core and reduction of clock frequency. The loss of the V_{dd} scaling handle, however, has been a significant setback to circuit designers. The root of the problem is inherent to the transistor itself, i.e. MOSFET.



$$I_{DS} \propto \exp\left(\frac{C_{ox}}{C_{ox}+C_{dep}}\frac{q}{kT}V_{GS}\right)$$

Figure 1.2: Principle of operation of MOSFET involves charge injection over a potential barrier. This limits the subthreshold swing to 60 mV/dec. at best.

Figure 1.2 shows the principle of operation of the MOSFET, where charge is injected over a gate controlled potential barrier. Since electrons in the source are distributed in energy by Boltzmann statistics, the current over the barrier has exponential kT dependence as shown. This suggests that at best kT/q decrease in gate voltage (assuming perfect gate control of barrier) results in a factor of e drop in current or 60 mV for 10X drop in current at room temperature. The MOSFET is inherently limited to 60 mV/dec. subthreshold swing. In order to scale supply voltage and maintain the same drive current or I_{on}, threshold voltage V_t must also be decreased seen from Eq. (1.2).

$$I_{on} \propto \left(V_{dd} - V_t\right)^{\alpha} \text{ where } \alpha = 1-2 \qquad (1.2)$$

However, reducing the V_t increases the I_{off} exponentially since the swing is limited to 60 mV/dec. Large I_{off} at low V_{dd} presents a significant standby power consumption problem, where power is consumed even when the chip is not performing computations. In addition, too low of an I_{on}/I_{off} ratio at low V_{dd} is a problem for sensitive circuits such as SRAM with regards to noise margin. To regain the supply voltage handle the swing must somehow be made less than 60 mV/dec.

1.2 Increasing Energy Usage in Server Farms

In addition to the integrated circuit thermal management issues discussed in Section 1.1, electricity usage of server farms and data centers is a growing concern in terms of emission, strain on the power grid, and cost to businesses. Internet use has become prevalent in our society as more and more information is being stored in the “cloud”. In 2006, the total US energy consumption from servers was estimated at 61 billion kWh with cost of $4.5 billion. This is equivalent to the electricity consumed by approximately 5.8 million U.S. households [1.2-1.4]. Figure 1.3 shows the projected annual electricity use of U.S. server farms assuming historical trends. By 2011, EPA estimates consumption of 100 billion kWh annually with cost of $7.4 billion. As an example, estimates of the electricity consumption of the entire country of Sweden are reported at 150 billion kWh, which is only 50% more than this projected 2011 value. [1.2-1.4] Figure 1.4 shows the breakdown of power allocation of a typical data center. Approximately 50% of energy use is from cooling and ventilation. This suggests significant improvement in data center efficiency can be realized by IC power reduction from a reduction in supply voltage V_{dd}. Unfortunately, the MOSFET 60 mV/dec. swing limits the effectives of this approach as described in section 1.1. A new “greener” transistor is desired that permits operation at low V_{dd} with acceptable performance.

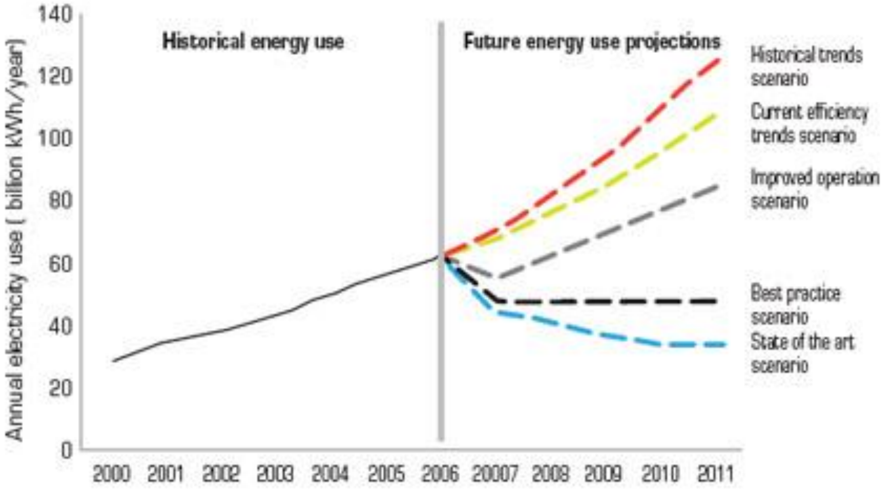


Figure 1.3: EPA projected trends of annual electricity use from U.S. server farms. [1.3]

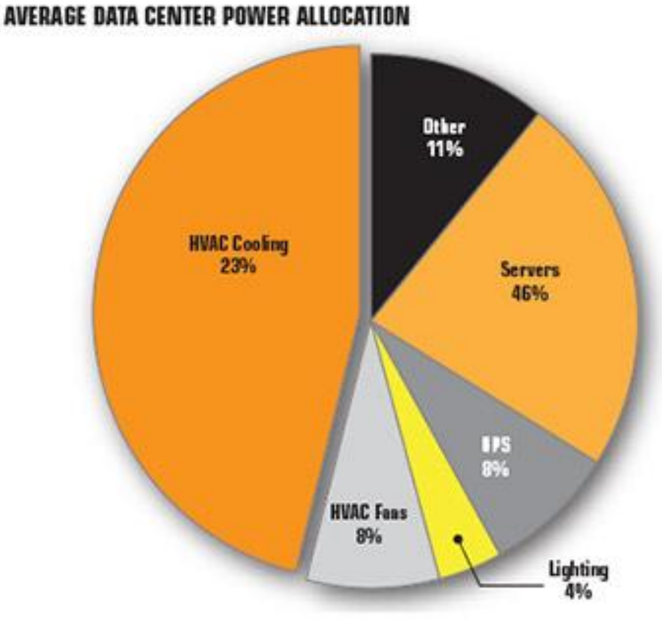


Figure 1.4: Breakdown of power allocation of typical data center. 50% of energy use goes to cooling. [1.3]

1.3 A Possible Solution

From Section 1.1, it seen that the MOSFET swing needs to be much less than 60 mV/dec. to regain the supply voltage scaling handle. This is fundamentally not possible with the MOSFET, since charge injection and Boltzmann statistics are involved. Tunneling or more specifically band-to-band tunneling, where electrons tunnel from the valance to conduction band of a semiconductor, is not theoretically subject to the 60 mV/dec. limit. Researchers have explored using band-to-band tunneling as a transistor “turn on” mechanism but with limited success in terms of simultaneous steep swing and drive current. [1.5-1.9] Newer tunnel transistor structures and designs need to be explored that may allow less than 60 mV/dec swing with good I_{on}.

1.4 Research Outline

The overall focus of this research is to understand the parameters needed for newer and “greener” solid state transistors with very steep subthreshold swing much less than 60 mV/dec over very large current ranges, permitting low supply voltage operation. The scope of this work is limited to transistors that operate via the tunneling process instead of thermal injection as in MOSFET. Chapter 2 reviews the tunneling principle and the derivation and approximations made for current band-to-band tunneling models. In Chapter 3, a tunneling field effect transistor (TFET) is introduced and an analytical frame work is developed. A newer TFET design called the “green” TFET or gTFET is proposed that allows for potential 200 mV supply voltage operation with appropriate band gap material from simulations. Chapter 4 reviews the experimental fabrication of the gTFET and measurement results. Chapter 5 introduces another steep swing tunneling based transistor using ultra thin silicon body (UTB) on silicon on insulator (SOI) substrate called UTB gTFET. Initial fabrication result of the UTB gTFET is also presented.

1.5 References

[1.1] ITRS Roadmap Executive Summary [online] Available: http://www.itrs.net/Links/2007ITRS/ExecSum2007.pdf

[1.2] J. G.Koomey, “Worldwide electricity used in data centers,” Environmental Research Letters. vol. 3, pp.1-8, 2008.

[1.3] R. Allen. “The greening of server farms.” Energy Efficiency and Technology. Sept. 2009. [online] Available: http://eetweb.com/applications/greening-server-farms-20091001/

[1.4] SG Equity Research, “Report to Congress on server and data center energy efficiency public law 109-431” US EPA, August 2, 2007. [online] Available: http://hightech.lbl.gov/documents/DATA_CENTERS/epa-datacenters.pdf

[1.5] W. M. Reddick, G. A. Amaratunga, "Silicon surface tunnel transistor," Applied Physics Letters, vol.67, no.4, pp.494-496, July 1995.

[1.6] C. Aydin, A. Zaslavsky, S. Luryi, S. Cristoloveanu, D. Mariolle, D. Fraboulet, S. Deleonibus, “Lateral interband tunneling transistor in silicon-on-insulator," Applied Physics Letters , vol.84, no.10, pp.1780-1782, March 2004.

[1.7] K. K. Bhuwalka, M. Born, M. Schindler, M. Schmidt, T. Sulima and I. Eisele, “P-channel tunnel field-effect transistors down to Sub-50 nm channel lengths”, Japanese Journal of Applied Physics,45, pp.3106-3109, 2006.

[1.8] V. Nagavarapu, R. Jhaveri, J.C.S. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," Electron Devices, IEEE Transactions on, vol.55, no.4, pp.1013-1019, April 2008.

[1.9] C. L. Royer and F. Mayer, “Exhaustive Experimental Study of Tunnel Field Effect Transistors from Materials to Architecture”, International Conference on Ultimate Integration of Silicon, pp.53-56, March 2009.

Chapter 2: Tunneling Phenomenon

2.1 Introduction

Quantum mechanical tunneling through potential energy barriers is a well understood phenomenon in the field of semiconductor devices. For instance, models for electron and hole tunneling through the gate insulator in MOSFET have been in excellent agreement with the experimental data. [2.1-2.4] This chapter reviews the most general tunneling framework and focuses in particular on the process of band-to-band tunneling (BTBT), in which electrons tunnel across the energy gap of a semiconductor (i.e., valance to conduction band). Subsequent chapters will discuss the simulation and experimental results of several transistor designs utilizing BTBT as an active "source" and enabler for ultra low voltage operation. A thorough understanding of the BTBT models is invaluable for future discussions. In this chapter the derivation of the local or constant electric field tunneling model is reviewed. It will be shown that the local model agrees well with a rigorous non-local tunnel probability calculation if an appropriate average electric field is used. Models for tunneling across hetero-structure interfaces are developed and compared to experimental measurement.

2.2 Original Kane Formulation

One of the original expressions for the rate of electrons tunneling from valance to conduction band (cm^{-3}s^{-1}) in a semiconductor was derived by Kane in his paper titled "Zener Tunneling in Semiconductors". [2.5] His basic approach used the concept of time dependant perturbation theory and Fermi's Golden Rule (shown in Eq. (2.1) to calculate the transition rate of carriers tunneling into the conduction band.

$$G_{btbt} = \frac{2\pi}{\hbar}\left|\langle \Psi_1 | H | \Psi_2 \rangle\right|^2 \tag{2.1}$$

The initial state $\langle \Psi_1 |$ can be described as a summation of Bloch states (eigenvalues of the periodic potential Hamiltonian) in the valance band, where as the final state $| \Psi_2 \rangle$ is a summation of conduction band Bloch states. H is the perturbation operator. The details of this matrix element calculation are very complex, subtle and lengthy. Kane describes theses mathematical details over 10 pages. For the purposes of this chapter only the end result is of interest shown in Eq. (2.2).

$$G_{btbt} = \frac{q^2\sqrt{m^*}E^2}{18\pi\hbar^2\sqrt{E_G}}\exp(-\frac{\pi\sqrt{m^*}E_G^{3/2}}{2\sqrt{2}q\hbar E}) \tag{2.2}$$

This equation shows that the functional form for the band-to-band tunneling rate has an exponential dependence on electric field. It should be noted that this basic functional form of $AE^2\exp(-B/E)$ is inherit to all tunneling phenomena, i.e. Schottky tunneling at metal/semiconductor interfaces or Fowler Nordheim and direct tunneling through gate insulators, each with different A and B coefficients. [2.6]

Eq. (2.2) was derived assuming various simplifications to allow for closed form. In the following sections in this chapter an alternative but more intuitive approach utilizing explicit calculation of the tunneling probability via the WKB approximation is detailed. The approach involves a summation over all valance band states with momentum directed towards the tunnel barrier appropriately weighted by a transmission probability. The most general expression will be shown to not be in closed form. However, when subjected to various approximations and simplifications, the final expression for tunneling rate is shown to be identical to Eq. (2.2) with the exception of slight difference in numerical pre-factor.

2.3 General Framework of Band-to-Band Tunneling

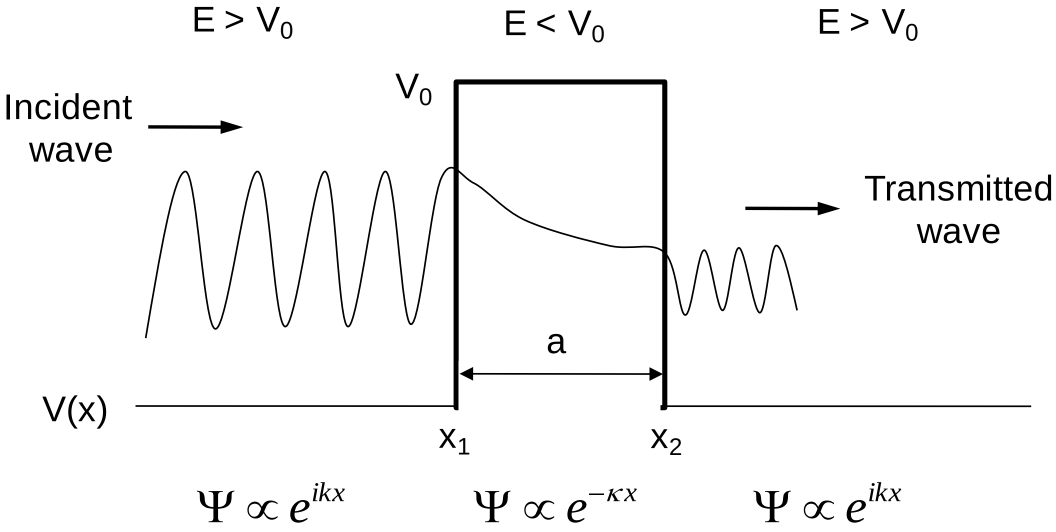


Figure 2.1: General tunneling problem setup for a rectangular potential barrier. The wave vector is real in the incident and transmitted regions and imaginary within the barrier. The imaginary wave vector leads to exponential decay of the wave function amplitude and decreased probability of transmission.

2.3.1 The WKB Approximation

The WKB approximation allows for the calculation of an approximate tunneling probability through arbitrary shaped potential barriers. A formal rigorous derivation is detailed in [2.7], however, in this section a heuristic argument is outlined to justify the WKB approach. Figure 2.1 shows the general tunneling problem setup for a rectangular potential barrier, where analytical solution to the Schrödinger equation exists in all three regions. By matching appropriate boundary conditions an expression for the tunnel probability (ratio of the transmitted to incoming wave amplitude) can be obtained as follows [2.7].

$$T = \frac{1}{1 + \frac{V_0^2}{4E\left(V_0 - E\right)} \sinh^2\left(\frac{a}{\hbar}\sqrt{2m\left(V_0 - E\right)}\right)} \tag{2.3}$$

For the case of a large or thick barrier or in general when the probability of tunneling is low (i.e., $\frac{a}{\hbar}\sqrt{2m\left(V_0 - E\right)} \gg 1$) Eq. (2.3) reduces to a much simpler form seen in Eq. (2.4).

$$T \cong \frac{4E(V_0 - E)}{V_0^2}\exp\left(-\frac{2a}{\hbar}\sqrt{2m\left(V_0 - E\right)}\right) \approx \exp\left(-\frac{2a}{\hbar}\sqrt{2m\left(V_0 - E\right)}\right) = \exp\left(-2\kappa a\right) \quad (2.4)$$

The pre-factor in Eq. (2.4) is on the order of unity and the expression is dominated by the exponential with final approximate form in the right hand side. Any arbitrary shaped barrier can be described as a series of rectangular barriers shown in Figure 2.2. The total transmission probability will be the product of the individual tunnel probabilities of each rectangular barrier.

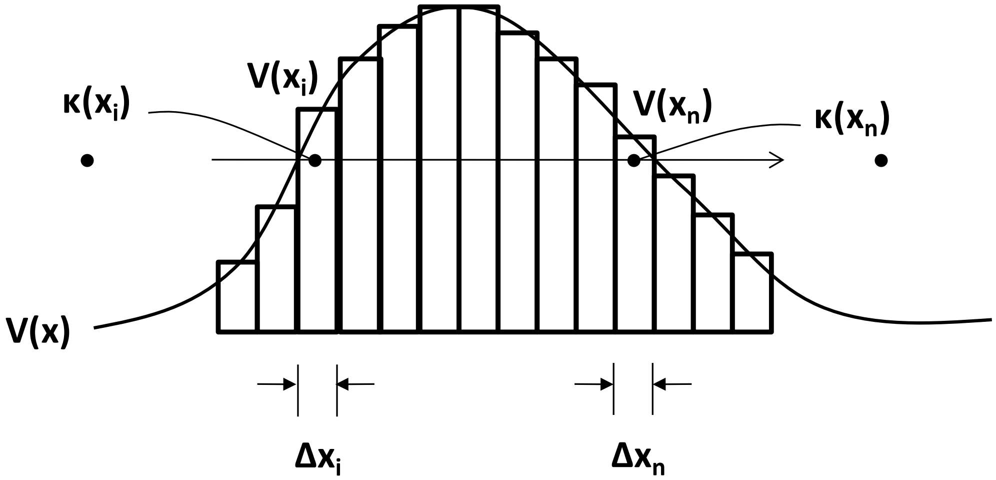


Figure 2.2: Any arbitrary potential barrier shape V(x) can be treated as a series rectangular barriers. The tunneling probability is the product of the tunneling probabilities through each rectangular barrier. The WKB approximation can be obtained in the limit as the barrier width approaches zero.

$$T \approx \prod_{i=1}^{n}\exp\left(-2\kappa_i \Delta x_i\right) = \exp\left(-2\sum_{i=1}^{n}\kappa_i \Delta x_i\right) \quad (2.5)$$

In the limit where the rectangular barrier thickness Δx_i is infinitesimally small, Eq. (2.6) is obtained, which is the definition of the WKB approximation.

$$T \approx \exp\left(-2\int_{x1}^{x2}\kappa(x)dx\right) \quad (2.6)$$

The classical turning points x_1 and x_2 are the locations where the electron enters and exits the potential barrier. The parameter which characterizes the amount of wave function decay within the barrier is the imaginary component of the wave vector *κ(x)*. To calculate the tunneling probability from the WKB approach an accurate expression of the imaginary wave vector throughout the electron path must be known.

2.3.2 Semiconductor Imaginary Wave Vector Dispersion Relation

For the process of band-to-band tunneling, where carries tunnel through the band gap of a semiconductor from valance to conduction band, it is difficult to visualize the shape of the potential barrier V(x). For the purposes of calculating the tunneling probability from the WKB approach Eq. (2.6), the exact shape or form of V(x) is not of concern but rather the imaginary wave vector relation *κ(x)* within the barrier (energy gap) that is important. More generally the imaginary wave vector dispersion relation *κ(E)* needs to be specified, which describes the amount of wave function decay as a function of the energy within the band gap relative to the valance band edge. *κ(x)* can be determined from *κ(E)* from knowledge of the electric field $\frac{d\phi}{dx}$ along the tunneling path as follows. (Note: E may refer to either energy or electric field depending on context)

$$\kappa(E) = \kappa(E = \int_0^x \frac{d\phi}{dx'} dx') = \kappa(x) \tag{2.7}$$

The simplest dispersion relation for the imaginary wave vector within the semiconductor band gap is the parabolic or 1-band relation. [2.6]

$$\kappa(E) = \frac{\sqrt{2m^*(E - E_V)}}{\hbar} \tag{2.8}$$

The imaginary wave vector at the valance band is zero, as is expected for a traveling electron wave impinging onto a barrier. Note however, that at the conduction band edge Eq. (2.8) predicts significant damping. In actuality, the transmitted electron is a traveling wave and should have zero imaginary wave vector component at the conduction band edge. To resolve this issue, a symmetric 2-band relation can be used [2.8-2.9].

$$\kappa(E) = \frac{\sqrt{2m^*}}{\hbar} \sqrt{\left(E - E_v\right)\left(1 - \frac{E - E_v}{E_g}\right)} \tag{2.9}$$

The imaginary component is zero at both band edges. Eq. (2.9) represents the simplest functional form that satisfies this property. However, only a single effective mass, m^* is used, which does not allow for the case where the valance and conduction band masses may be different. Eq. (2.10) is the non-symmetric 2-band relation taking into account differing effective mass [2.10-2.12].

$$\begin{gathered}\kappa_v(E) = \frac{\sqrt{2m_v^*}}{\hbar}\sqrt{\left(E - E_v\right)} \text{ and } \kappa_c(E) = \frac{\sqrt{2m_c^*}}{\hbar}\sqrt{\left(E_c - E\right)} \\ \kappa(E) = \min\left(\kappa_v(E), \kappa_c(E)\right)\end{gathered} \tag{2.10}$$

Figure 2.3 plots all three dispersion relations using silicon parameters (m_c =0.2 $m_{v,hh}$ = 0.44). These values are for [100] tunneling from the heavy hole band to the conduction band transverse axis ellipsoidal. The non-symmetric 2-band relation has been shown to be in very good agreement with calculated values of the complex band structure within the band gap using pseudo-potential methods for silicon. [2.12] The band-to-band tunneling probability can therefore be calculated from the WKB approach of Eq. (2.6) using the imaginary component of wave vector relation from Eq. (2.10).

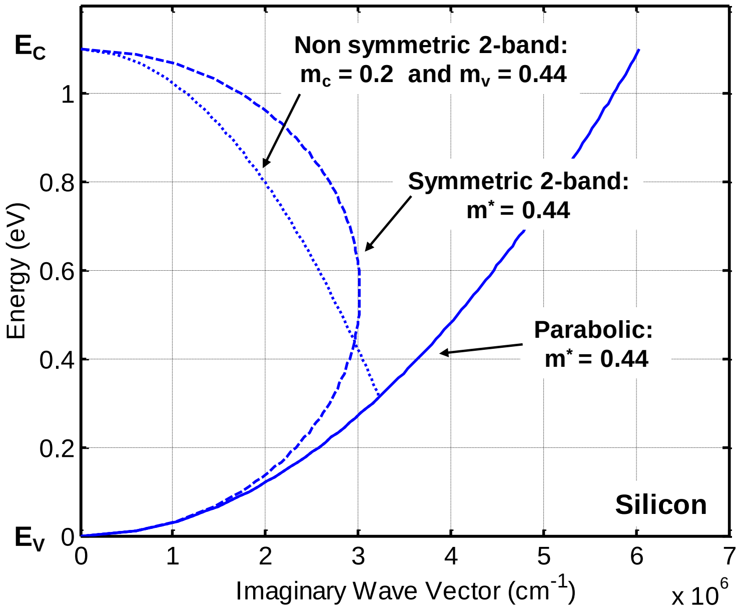


Figure 2.3: Various imaginary wave vector dispersion relations within the band gap are plotted for silicon. Non-symmetric 2-band has been shown to be the most rigorously correct. [2.12]

2.3.3 Complications of 3D Tunneling

In a bulk semiconductor in a three dimensional system, additional complexities to the tunneling problem arise. Because of the extra dimensions, it is possible for the tunneling electron to have part of its total momentum directed transverse to the tunneling direction. This transverse energy must be taken into account in the tunneling calculation. The imaginary wave vector relation needs to be modified as follows (assuming *x* is the direction of tunneling) [2.13].

$$E_v - E = \frac{\hbar k_x^2}{2m_v^*} + \frac{\hbar k_y^2}{2m_v^*} + \frac{\hbar k_z^2}{2m_v^*} = \frac{\hbar k_x^2}{2m_v^*} + E_T \text{ and } E - E_c = \frac{\hbar k_x^2}{2m_c^*} + \frac{\hbar k_y^2}{2m_c^*} + \frac{\hbar k_z^2}{2m_c^*} = \frac{\hbar k_x^2}{2m_c^*} + E_T$$

$$\kappa_v(E,E_T) = \frac{\sqrt{2m_v^*}}{\hbar}\sqrt{\left(E - E_v\right) + E_T} \text{ and } \kappa_c(E,E_T) = \frac{\sqrt{2m_c^*}}{\hbar}\sqrt{\left(E_c - E\right) + E_T} \qquad (2.11)$$

$$\kappa(E,E_T) = \min\left(\kappa_v(E,E_T), \kappa_c(E,E_T)\right)$$

Note that with increasing non-zero transverse energy the amount of wave function damping and therefore tunneling probability is decreased. Conservation of momentum also implies that the transverse energy must be conserved across the tunnel barrier. This has implications on the classical turning points as shown in Figure 2.4. For a given E_T the barrier width is increased and $x_{1'}$ and $x_{2'}$ are such that Eq. (2.11) is zero. The turning points will change for different values of transverse energy.

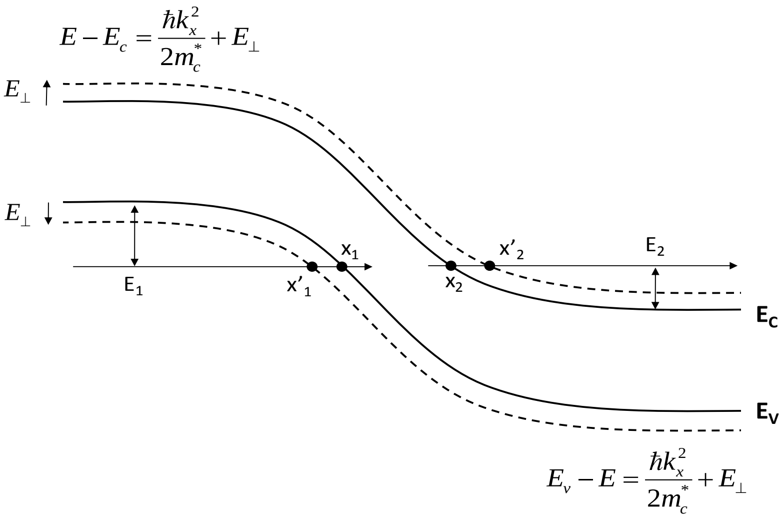

$E-E_c=\frac{\hbar k_x^2}{2m_c^*}+E_\perp$
$E_\perp$
$E_\perp$
x₁
x'₂
E₂
E₁
x'₁
x₂
E_C
E_V
$E_v-E=\frac{\hbar k_x^2}{2m_c^*}+E_\perp$

Figure 2.4: An example energy band diagram demonstrating process of band-to-band tunneling. x_1 and x_2 are the classical turning points for the tunneling path indicated in the arrow. For non-zero transverse energy E_T the barrier width increases with new turning points x_1' and x_2'.

2.3.4 Derivation of the Band-to-Band Tunneling Current

The PN junction energy band diagram of Figure 2.4 is used as an example in this derivation. To calculate the band to band tunneling current for Figure 2.4 the number of states in *k*-space in volume $dk_xdk_ydk_z$ around (k_x,k_y,k_z) and $(k_x+dk_x,\ k_y+dk_y,\ k_z+dk_z)$ is calculated in the valance band of the emitting side. [2.13-2.14]

$$dn=\frac{2}{(2\pi)^3}dk_xdk_ydk_z \tag{2.12}$$

The constant pre-factor is the three-dimensional density of states in *k*-space. The current in the direction of the barrier from this differential *k*-space volume is determined as follows.

$$dJ_x=q\frac{2}{(2\pi)^3}v_xdk_xdk_ydk_z=q\frac{2}{(2\pi)^3}\frac{1}{\hbar}\frac{\partial E}{\partial k_x}dk_xdk_ydk_z=q\frac{2}{(2\pi)^3}\frac{1}{\hbar}dEdk_ydk_z \tag{2.13}$$

The velocity towards the barrier can be expressed as the group velocity of valance band electron wave packet $v_x=\frac{1}{\hbar}\frac{\partial E}{\partial k_x}$. The fraction of the current that tunnels through the barrier is given by the tunneling probability $T(E,E_T)$ of Eq. (2.6) taking into account transverse energy E_T. The band-to-band tunneling current from this volume in *k*-space can be expressed as follows.

$$dJ_{btbt} = q\frac{2}{(2\pi)^3}\frac{1}{\hbar}T(E,E_T)\,dE\,dk_y dk_z \tag{2.14}$$

The area element *dk_ydk_z* can also be expressed in terms of dk_t, the transverse wave vector.

$$\begin{gathered} k_t^2 = k_y^2 + k_z^2 \text{ and } dk_y dk_z = 2\pi k_t dk_t \\ dJ_{btbt} = q\frac{2}{(2\pi)^3}\frac{1}{\hbar}T(E,E_T)\,dE\,2\pi k_t dk_t \end{gathered} \tag{2.15}$$

The variables can then be changed in terms of the transverse energy dE_T as follows.

$$\begin{gathered} E_T = \frac{\hbar^2 k_t^2}{2m} \text{ and } \mathrm{dE}_T = \frac{\hbar^2}{m}k_t dk_t \\ dJ_{btbt} = q\frac{2}{(2\pi)^2}\frac{m}{\hbar^3}T(E,E_\perp)\,dE\,dE_T = \frac{4\pi qm}{h^3}T(E,E_\perp)\,dE\,dE_T \end{gathered} \tag{2.16}$$

To calculate the current density Eq. (2.16) is integrated over the energy *E* and transverse energy E_T. The integration on *E* is performed over the entire overlap between valance and conduction band. The limit of integration of the E_T is the min(E_1,E_2) and depends on the energy *E* as shown in Figure 2.4.

$$J_{btbt} = \frac{4\pi qm}{h^3}\int_0^E \int_0^{\min(E_1,E_2)} T(E,E_\perp)\,dE\,dE_T \tag{2.17}$$

To derive the generation rate a change of coordinates from *E* to position *x* is performed.

$$G_{btbt} = \frac{4\pi qm}{h^3}\frac{dE}{dx}\int_0^{\min(E_1,E_2)} T(E,E_\perp)\,dE_T \tag{2.18}$$

2.3.5 The Local Approximation

Eq. (2.18) is the expression for non-local band-to-band generation rate. However, it is not in closed form since it contains an integral over E_T and the expression for the tunneling probability is itself an integral. In order to obtain a closed form expression for the generation rate a series of approximation must be made. (1) The imaginary wave vector expression is taken be the symmetric 2-band relation of Eq. (2.9). (2) The electrical field is assumed constant across the tunneling path. These two approximations permit the tunneling probability to be calculated in closed form as follows.

$$T\left(E_T\right)=\exp\left(-\frac{\pi}{2\sqrt{2}}\frac{m^{1/2}E_g^{3/2}}{q\hbar E}\right)\exp\left(-E_T\,\frac{\pi\left(2mE_g\right)^{1/2}}{q\hbar E}\right)\text{ where E is electric field}\qquad(2.19)$$

The integration limit on E_T from Eq. (2.18) is also taken to be infinity. This is typically justified because the tunneling probability decreases exponentially with E_T. The final result is given as follows.

$$G_{btbt}=\frac{q^2\sqrt{m}}{2\pi^3\sqrt{2}\hbar^2\sqrt{E_g}}E^2\exp\left(-\frac{\pi}{2\sqrt{2}}\frac{m^{1/2}E_g^{3/2}}{q\hbar E}\right)=AE^2\exp\left(-\frac{B}{E}\right)\qquad(2.20)$$

Compared to Eq. (2.2) derived from Kane's paper using Fermi's golden rule, Eq. (2.20) is nearly identical with the exception of the pre-factor, which is approximately $\sqrt{2}$ times larger. The coefficients *A* and *B* are the band-to-band tunneling parameters of the local tunneling model. For silicon these parameters have been calibrated to experimental data.[2.15] Eq. (2.20) is sometimes referred to as the local model and is the band-to-band tunneling model used in almost all device simulators.

2.3.6 Validity of the Local Approximation

In order to obtain a closed form expression for the tunneling probability two major approximations needed to be made. (1) The imaginary wave vector dispersion relation was taken to be the 2-band symmetric form. (2) The electric field across the tunneling path was assumed constant. The latter is sometimes referred to as the local approximation.

$$T_{non-local}\approx\exp\left(-2\int_{x1}^{x2}\frac{\sqrt{2m^*}}{\hbar}\sqrt{(qE_g-q\int_0^x E(x^{'})dx^{'})(1-\frac{(E_g-\int_0^x E(x^{'})dx^{'})}{E_g})}dx\right)$$

$$\text{when E-field E(x) is constant: }T_{local}\approx\exp\left(-\frac{\pi}{2\sqrt{2}}\frac{\sqrt{m^*}E_g^{3/2}}{e\hbar E}\right)\qquad(2.21)$$

In actuality, the field across the tunneling path is almost never constant. For example, in regions of constant doping the electric field varies linearly. This brings into the question the validity of the local approximation since the tunneling process is inherently very non-local. Since many of the numerical tunneling models employed in commercial semiconductor device simulation tools assume this approximation, some caution must be taken when using these models.

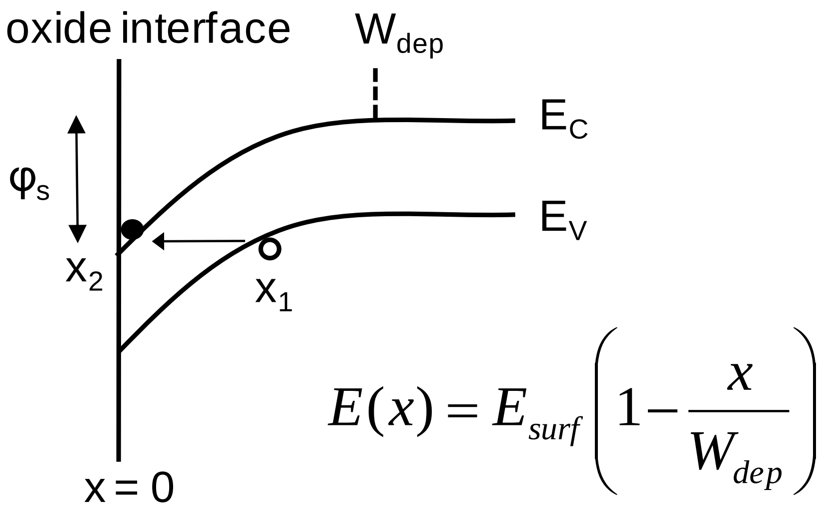


Figure 2.5: A band-to-band tunneling scenario commonly seen in the drain overlap region of MOSFET in "off-state". The electric field E(x) varies linearly across the tunneling path and is not constant.

Figure 2.5 shows the band bending profile seen in the drain overlap region of a MOSFET when biased in its "off-state". The cutline is made normal to the gate dielectric interface. Band-to-band tunneling can occur in this overlap region (known as gate induced drain leakage, GIDL [2.16]). To first order the region can be treated as one of constant doping concentration. This leads to a linear electric field variation across the tunnel path. This, however, leads to some ambiguity on which value of field E to choose when using the local model of Eq. (2.21) for calculation of the tunneling probability, since the field is not constant.

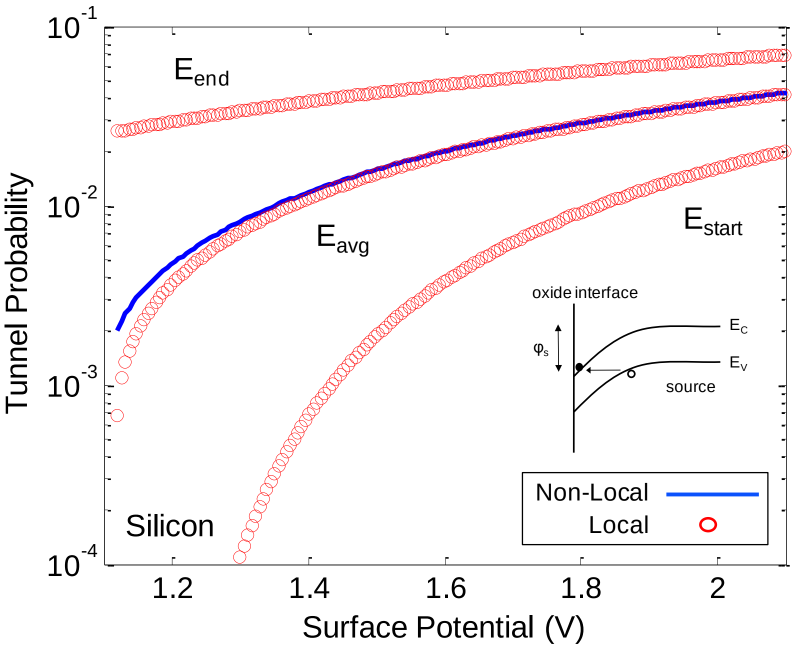


Figure 2.6: Tunneling probability is calculated as function of total band bending in the silicon for the situation outlined in the lower right subfigure. The path which ends at the oxide interface is calculated. When using an average electric field across the path in the local model the result agrees reasonably well with the non-local calculation.

In Figure 2.6 the tunneling probability is calculated for the situation outlined in Figure 2.5 (linearly varying electric field) as function of the surface potential (or total band bending). The non-local calculation is shown with the solid line. If the field at the start of the path is used in the local model, the result greatly underestimates the actual tunnel probability. If the field value at the end of the path is used the result significantly overestimates. Using an average electric field value across the total path (i.e. the value on average which the electron “experiences”) agrees reasonably well with the non-local calculation. This result justifies the local approximation as long as the proper field calculation is employed. This is significant because most simulation tools employ the local tunneling model in some form. In addition, closed form expressions are greatly preferred when developing simple analytical models for tunneling current as will be detailed in Chapter 3.

2.3.7 Comparison of Local Model with Experimental Data

The previous section showed that the local tunneling probability (or constant electric field model) is a good approximation to the actual non-local tunneling problem if average electric field is used. The TCAD device simulator MEDICI implements the local tunneling model of Eq. (2.20) and has option on choice of electric field value. In this section, the local model with average electric field is compared with experimental tunneling current data from literature. In [2.17] measurement data from a tunnel field effect transistor (TFET) is presented. The TFET, which will be discussed in significant detail in Chapter 3, is a device where voltage on a gate terminal can modulate band-to-band tunneling current. Doping profile information, physical gate length, and oxide thickness of the TFET is detailed in [2.17] and implemented as accurately as possible in the device simulation. Figure 2.7 shows the comparison of the simulation results of the local tunneling model with average electric field with the experimental data using default A, B calibrated silicon tunneling parameters.

$$\begin{aligned} A &= 3.5\cdot 10^{21}\ \mathrm{cm^{-1}s^{-1}V^{-2}} \\ B &= 22.5\cdot 10^{6}\ \mathrm{V/cm} \end{aligned} \tag{2.22}$$

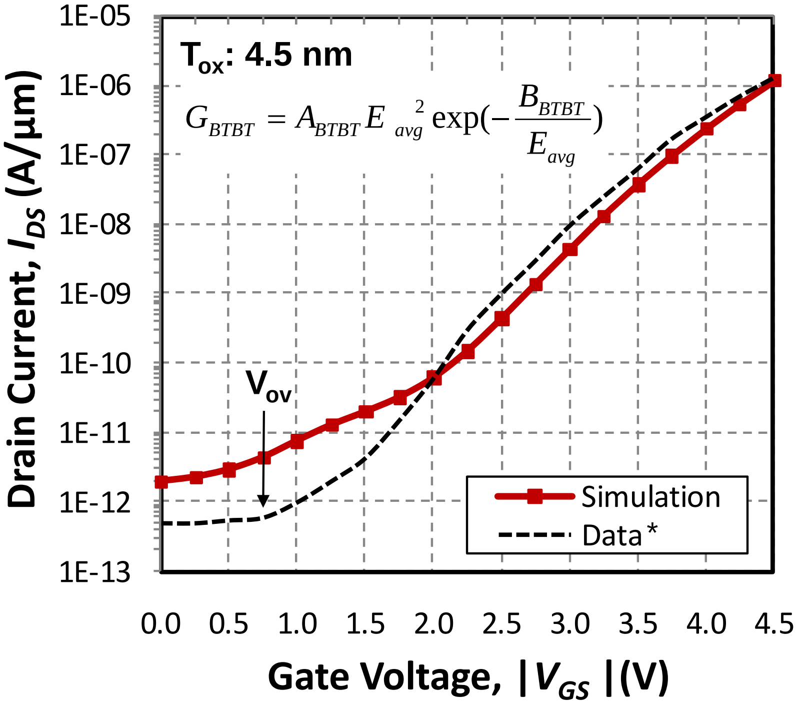


Figure 2.7: Comparison of local band-to-band tunneling model with average electric field with TFET experimental data [2.17]. Identical structure as that in [2.17] is simulated.

Reasonable agreement is seen with the experimental data for the average field local tunneling model in silicon. This gives confidence that this model is well suited for detailed simulation study to be presented in Chapter 3.

2.4 Hetero-Structure Tunneling Models and Effective Band Gap

So far the models presented in the previous chapter involve tunneling from the valance band of one material to the conduction band of that same material. An advantageous situation may arise where tunneling across the interface between two different semiconductors, i.e. from valance band of one material to the conduction band of another material. In particular, for a type II hetero-structure band offset, shown in Figure 2.8, the effective energy gap for tunneling can be smaller than that of either material. The effective energy gap depends on the band gap of the first semiconductor and the amount of conduction band offset ΔE_c of the second material. Non-local hetero-tunneling models are developed and implemented in MATLAB in this section and the concept of effective energy gap or $E_{g,eff}$ is discussed.

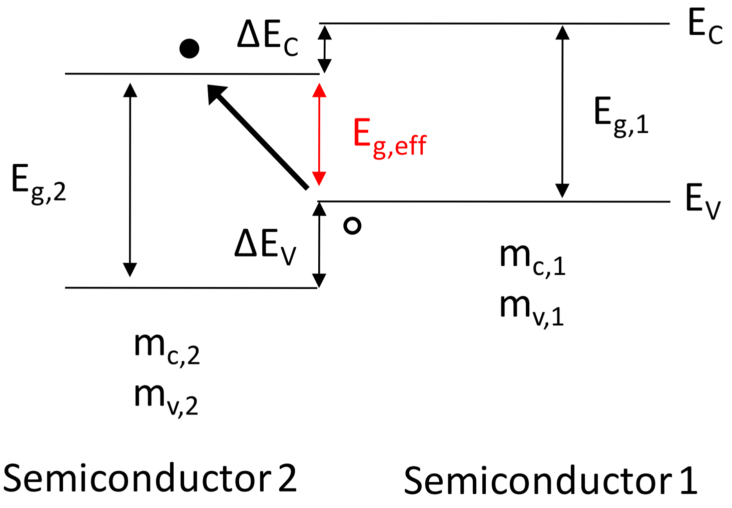


Figure 2.8: A type-II hetero-structure band offset permits tunneling across the effective band gap, $E_{g,eff}$. The $E_{g,eff}$ is smaller than the band gap of either semiconductor. Larger conduction band offset permits smaller $E_{g,eff}$.

2.4.1 A Non-Local Hetero-Tunneling Model

Unlike the case for tunneling in a single material, it is not possible to develop a simplified closed form expression for tunneling across the hetero-interface between two semiconductors by assuming constant electric field. Depending on the extent of overlap between energy bands the tunneling path may be entirely in one material or traversing through both materials as seen from Figure 2.9. The case where the electron travels through both materials corresponds to the smallest effective energy gap $E_{g,eff}$ and largest tunneling probability. For this case, the valance and conduction band effective mass and energy gap of both materials needs to be taken into account. In general, the tunneling current must be calculated non-locally by slight modification of the WKB framework developed in the previous section over the entire energy band overlap. All tunneling paths each with possibly different $E_{g,eff}$ must be considered.

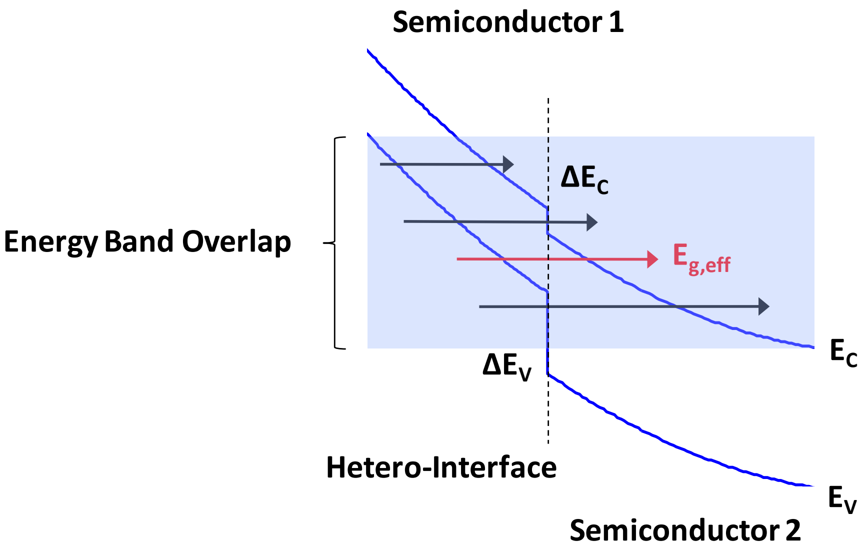


Figure 2.9: Various tunneling paths exist in a type-II band offset in addition to the one with energy gap of *E_{g,eff}*. (shown in red) A proper hetero-tunneling model needs to account for all paths in the shaded overlap region.

It will prove to be more useful to start with the tunneling current density Eq. (2.17). The effective mass in the pre-factor is the valance band effective mass of the emitting side (semiconductor 1). The calculation of the tunneling probability in the integrand needs to be generalized to take into account the presence of a hetero-structure. The imaginary wave vector dispersion relation and consequently tunneling probability are modified as follows.

$$\kappa_v(E(x),E_T)=\frac{\sqrt{2m_v^*(x)}}{\hbar}\sqrt{\left(E(x)-E_v(x)\right)+E_T} \text{ and } \kappa_c(E(x),E_T)=\frac{\sqrt{2m_c^*(x)}}{\hbar}\sqrt{\left(E_c(x)-E(x)\right)+E_T}$$

$$\kappa(E(x),E_T)=\min\left(\kappa_v(E(x),E_T),\kappa_c(E(x),E_T)\right)$$

$$T\approx\exp\left(-2\int_{x1'}^{x2'}\kappa(E(x),E_T)dx\right)$$

(2.23)

The effective masses are taken to be position dependant to allow for the case of tunneling across the hetero-structure, where mass changes along the tunneling path. This means that the overall tunneling probability is the product of the tunneling probability through semiconductor 1 and 2. The band bending profile $E_c(x)$ and $E_V(x)$ with included band offset must be known to calculate the energy within the forbidden gap for a given tunneling path. Eq. (2.17) with modified tunneling probability Eq. (2.23) is a complicated triple integration problem. For a particular path, numerical integration needs to be performed to calculate the tunneling probability. This is then integrated over the transverse energy limits discussed in section 2.3.3. Finally, integration is performed over the entire band overlap region taking into account all possible tunneling paths. This hetero-tunneling model has been implemented in a MATLAB program utilizing advanced built in numerical integration functions, such as quad() (for quadrature integration).

The band bending profile with position dependant parameters $E_c(x)$, $E_v(x)$, $m_c(x)$, and $m_v(x)$ are treated as the input to the model. This can be the output from a numerical device simulator for a particular voltage bias or from simple analytical solutions of Poisson equation. The energy band overlap is identified (i.e., shaded blue region in Figure 2.9). For a given tunneling path within the overlap region and for a finite transverse energy the classical turning points x_1' and x_2' are computed from the input $E_c(x)$ and $E_v(x)$ profile. Note that the turning points will change with transverse energy as discussed in section 2.3.3. The tunneling probability is then numerically integrated over the transverse energy limits and over the entire energy band overlap range. Use of MATLAB vectorization techniques helps with this computationally intensive calculation.

2.4.2 Concept of Effective Band Gap

The concept of effective band gap can be seen very clearly from plots of the imaginary wave vector along the tunneling direction. This discussion corresponds to the situation of band-to-band tunneling occurring normal to the gate dielectric in the MOSFET drain overlap region as was described in Figure 2.5. This type of vertical tunneling controlled by a gate voltage will be shown to be very important in Chapter 3 and the remainder of this thesis. For simplicity, effective masses are assumed identical across the hetero-structure to demonstrate the $\mathrm{E}_{\mathrm{g,eff}}$ concept. An arbitrary thin 1 nm semiconductor of E_g = 1.1 eV is placed atop another arbitrary semiconductor with E_g = 0.67 eV with arbitrary conduction band offset ΔE_c of 0.2 eV. This corresponds to an effective band gap of 0.47 eV. Figure 2.10 shows the imaginary wave vector curves at the initial point of tunneling overlap for three structures: (1) single semiconductor of E_g = 0.67 eV (2) hetero-structure described above (3) an arbitrary semiconductor with $E_g = E_{g,eff}$ = 0.47 eV. A larger tunneling probability corresponds to a smaller area under the wave vector curve. As shown, hetero-structure tunneling corresponds approximately with that of a semiconductor of $E_g = E_{g,eff}$.

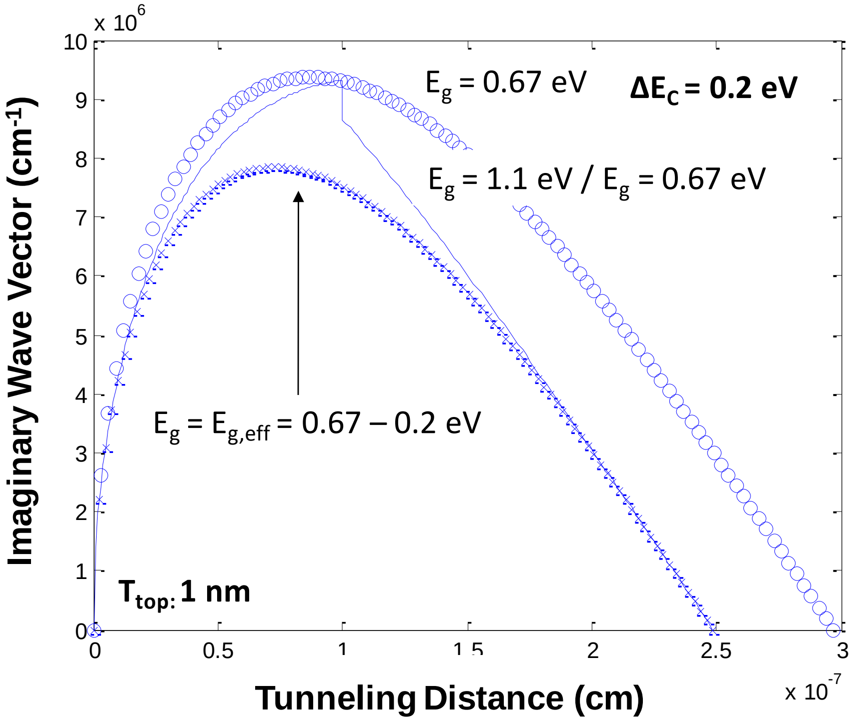


Figure 2.10: Imaginary wave vector along tunneling path for three structures. (1) Semiconductor of E_g = 0.67 eV (2) Hetero-structure of 1nm E_g = 1.1 eV atop E_g = 0.67 eV with conduction band offset of 0.2 eV. (3) Semiconductor of E_g = $E_{g,eff}$ = 0.67 – 0.2 = 0.4 eV. Tunneling across the hetero-structure is similar to tunneling in an $E_{g,eff}$ material.

2.4.3 s-Si/Ge Hetero-Junction Diode Example

The reverse tunneling current from a hetero-junction diode is simulated using the hetero-tunneling model developed in this section. In this case, a strained silicon / relaxed germanium diode (100) is used as an example. The lattice mismatch between silicon and germanium results in significant strain in the silicon, causing a large conduction band offset and ultra low $E_{g,eff}$. [2.18-2.20] In this case, the strained silicon and germanium layers are thick enough such that the depletion region is contained within the boundaries of the structure. Note that in actuality, beyond a thickness of approximately 1-2 nm silicon loses all of its strain when lattice matched to germanium. [2.20] This ideal structure serves largely to demonstrate the hetero-tunnel concept and model. As a control comparison a pure germanium diode is also simulated using the same model. The doping concentration is 5E19 cm^{-3} on both sides of the junction. The parameters used in the simulation are shown in the table below. [2.18-2.20]

	m_c	m_v	E_g	ΔE_c
Strained-Si	1.08	0.16	0.41 eV	0.45 eV
Germanium	0.12	0.044	0.67 eV	

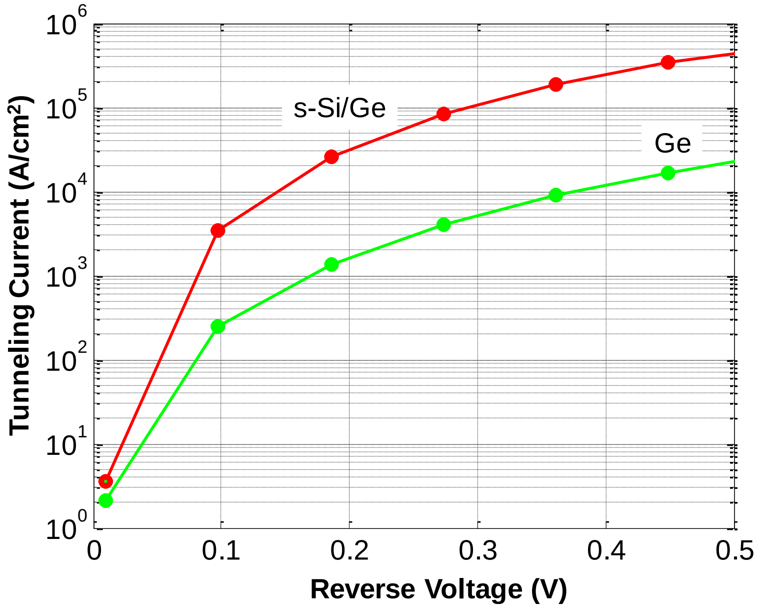


Figure 2.11: Simulated hetero-junction diode reverse tunneling current using hetero-tunneling model. The doping was 5E19 cm^{-3} on both sides of the junction. Relevant parameters are shown in the table on previous page. The s-Si/Ge diode shows larger tunneling current from the smaller $E_{g,eff}$ compared to Ge.

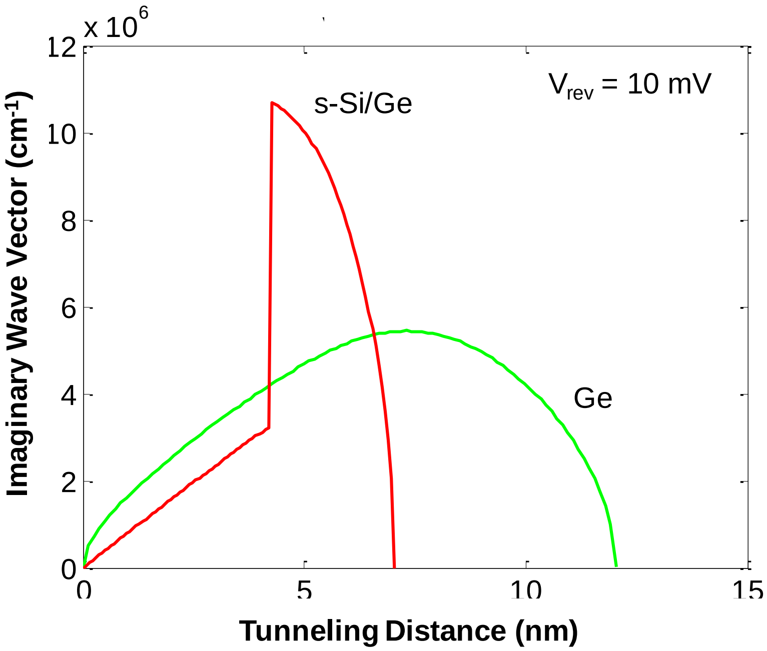


Figure 2.12: Imaginary wave vector along tunneling path for the hetero-junction diode at V_{rev} = 10 mV. The larger m_c of strained silicon results in large k vector increase when electron enters the silicon.

Figure 2.11 shows the simulated reverse bias tunneling current for the hetero-junction diode and germanium control diode. Tunneling current is significantly increased for the hetero-junction diode especially at the larger reverse bias. The effective band gap $E_{g,eff}$ is approximately 0.22 eV for this case. However, at small reverse bias the advantage of the hetero-structure is minimal. This can be explained by the large conduction band effective mass of strained silicon, which is caused by the lowering of the Δ_2 (longitudinal mass) band. Figure 2.12 shows the imaginary wave vector along the tunneling path at small reverse bias. The large increase in wave vector value corresponds to the transition point between germanium and strained silicon with the large m_c. This results in significant wave function decay and consequently lower tunneling probability even with $E_{g,eff}$ of 0.22 eV.

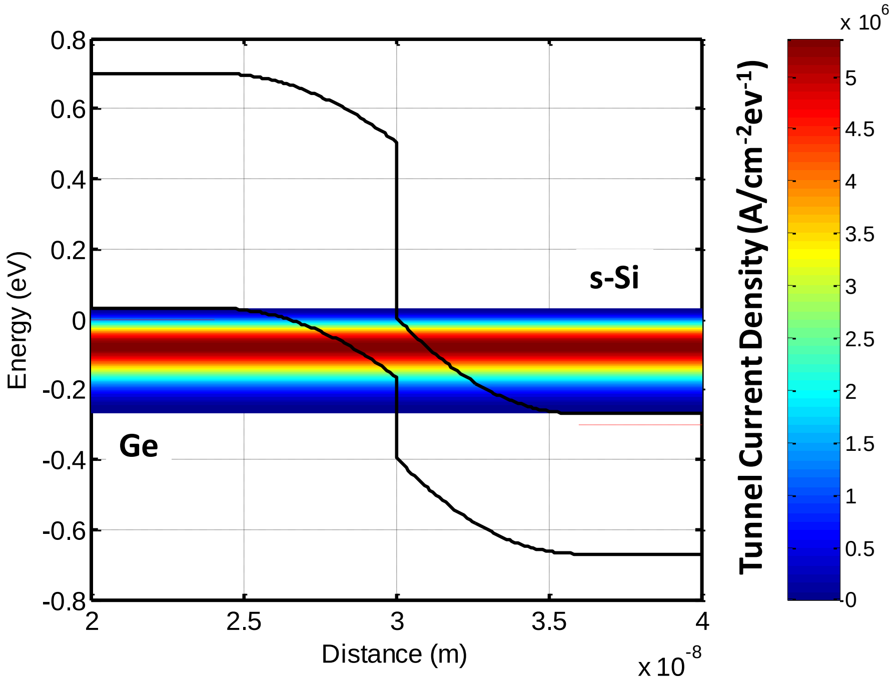


Figure 2.13: Simulation output of the hetero-tunneling model showing tunneling current density in units of A/cm^{2}eV for s-Si/Ge diode. The peak current flows through the hetero-structure with lowest $E_{g,eff}$ as expected.

Figure 2.13 shows the spectral current density (A cm^{-2} eV^{-1}) output from the hetero-tunneling model for the s-Si/Ge diode at reverse bias of 300 mV. Current largely flows through the hetero-structure where effective band gap is lowest as expected.

2.4.4 Initial Comparison of Hetero-Tunneling Model with Data

In this section the hetero-tunneling model is compared with hetero-junction diode data found in literature. The electrical characteristics of a lattice matched GaSb/AlSb/InAs hetero-structure diode are presented in [2.21]. These diodes are grown by molecular beam epitaxy with AlSb thickness of 2 nm. The InAs is doped to N_D = 7E17 cm^{-3} while the AlSb is intrinsic. The GaSb is doped p-type with doping level of N_A = 4E19 cm^{-3} treated as a fitting parameter. The effective masses, band gap, and band offsets used for these materials are those determined by researchers. [2.22] A resistance of 4 Ω is included to account for series resistance. Figure 2.14 shows the band diagram and comparison of the hetero-tunneling model with the experimental data for forward bias. (diode 2114 in [2.21]) Reasonable agreement is seen up until the peak current. This gives some initial confidence on the correctness of the developed hetero-tunneling model. More experimental comparison is needed in future work.

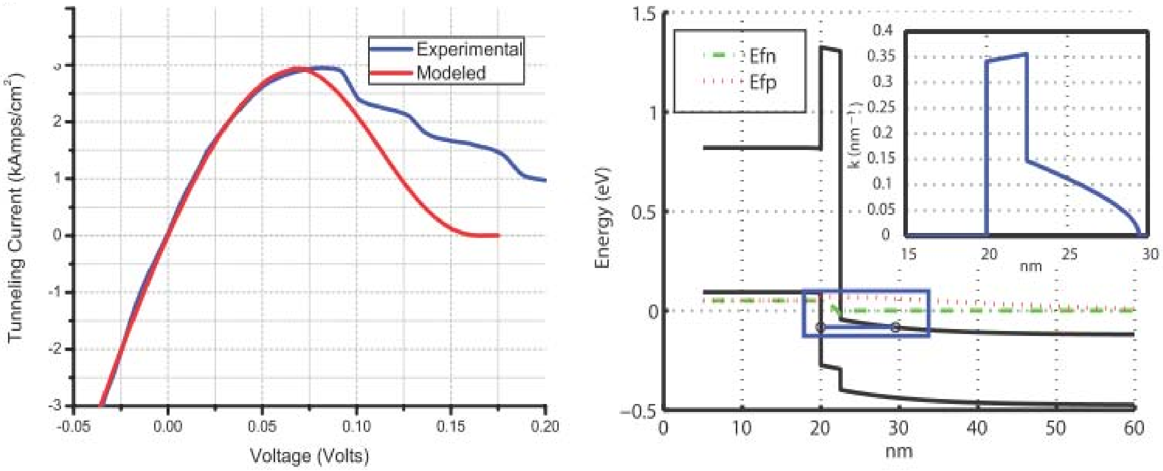


Figure 2.14: Comparison of GaSb/AlSb/InAs hetero-structure diode with implemented hetero-tunneling model. Reasonable agreement with experimental data [2.21] is seen.

2.5 References

[2.1] R. Stratton, “Theory of field emission from semiconductors,” Physical Review, vol.125, no.1, pp. 67-82, Jan 1962.

[2.2] J. Maserjian and G. P. Peterson, “Tunneling through thin MOS structures: dependence on energy (E-k),” Applied Physics Letters, vol.25, no.1, pp.50-52, July, 1974.

[2.3] G. Krieger and R. M. Swanson, “Electron tunneling in Si-SiO$_2$-Al structures: A comparison between (100) oriented and (111) oriented Si,” Applied Physics Letters, vol.39, no.10, pp.818-819, Nov 1981.

[2.4] C. Chang, “Tunneling in thin gate oxide MOS structures,” Ph.D Dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 1984.

[2.5] E. O. Kane, “Zener tunneling in semiconductors,” J. Physics Chemical Solids, vol.12, pp.181-188, 1959.

[2.6] S. M. Sze, Physics of Semiconductor Devices: 2nd Edition, John Wiley & Sons, New York, 1981.

[2.7] D. J. Griffiths, Introduction to Quantum Mechanics: 2nd Edition, Prentice Hall, pp.315, Upper Saddle River, 2005.

[2.8] Charles Kittle, Introduction to Solid State Physics, 4th Edition, p. 317, Jon Wiley & Sons, New York, 1971.

[2.9] G. W. Lewicki and C. A. Mead, Physics Rev Letters, vol.16, pp.939, 1969.

[2.10] P.V. Dressendorfer, “Interface and electron tunneling properties of thin oxides on silicon,” Ph.D Dissertation, Department of Engineering and Applied Science, Yale University, 1978.

[2.11] H. Flietner, “The E(k) relation for a two-band scheme of semiconductors and the application to the metal-semiconductor contact,” Physica Status Solidi (b), vol. 54 pp.201-208, 1972.

[2.12] M.V. Fischetti, T.P. O’Regan, S. Narayanan, C. Sachs, J. Seonghoon, J. Kim, Y. Zhang, “Theoretical study of some physical aspects of electronic transport in nMOSFETs at the 10-nm gate-length,” Transactions on Electron Devices, vol. 54, no.9, pp.2116-2136, Sept. 2007.

[2.13] S. Wang, Fundamentals of Semiconductor Theory and Device Physics, Prentice-Hall, 1989, pp.484-491.

[2.14] P. J. Price and J.M. Radcliffe, “Esaki Tunneling,” IBM Journal, pp.364-371,October 1959.

[2.15] H. J. Wann, P.K. Ko, C. Hu, “Gate-induced band-to-band tunneling leakage current in LDD MOSFETs,” International Electron Device Meeting, pp.6.5.1-6.5.4, 1992.

[2.16] T. Y. Chan, J. Chen, P. Ko, C. Hu, “The impact of gate-induced drain leakage current on MOSFET scaling”, International Electron Devices Meeting, Vol.33, pp: 718-721, 1987.

[2.17] K. K. Bhuwalka, M. Born, M. Schindler, M. Schmidt, T. Sulima and I. Eisele, “P-channel tunnel field-effect transistors down to Sub-50 nm channel lengths”, Japanese Journal of Applied Physics,45, pp.3106-3109, 2006.

[2.18] C. G. Van der Walle and R. M. Martin, “Theoretical calculations of heterojunction discontinuities in the Si/Ge System,” Physics Review B, vol. 38(8), pp.5621-5634, 1986.

[2.19] M. M. Rieger and P. Vogl, “Electronic-band parameters in strained SiGe alloys on SiGe substrates,” Physics Review B, vol. 48(19), pp.14276-14287, 1993.

[2.20] R. Delhougne, G. Eneman, M. Caymax, R. Loo, P. Meunier-Beillard, P. Verheyen, W. Vandervorst, K. De Meyer, M. Heyns, “Selective epitaxial deposition of strained silicon: a simple and effective method for fabricating high performance MOSFET devices,” Solid-State Electronics, vol.48, no.8, pp.1307-1316, August 2004.

[2.21] J. Schulman and D. Chow, “Sb-Heterostructure interband backward diodes,” Electron Device Letters, vol. 21, no. 7, pp. 353–355, 2000.

[2.22] S. Tiwari and D. J. Frank, “Empirical fit to band discontinuities and barrier heights in III–V alloy systems,” Applied Physics Letters, vol. 60, no. 5, pp. 630–632, Feb 1992.

Chapter 3: A “Green” Tunnel Field Effect Transistor

3.1 Introduction

The principle of operation of a transistor that is theoretically capable of achieving sub 60 mV/dec is explored. The tunnel field effect transistor or TFET uses band-to-band tunneling, a process not subject to the 60 mV/dec limit, as an on-state mechanism. Most researchers have explored the basic gated PN diode TFET design [3.1-3.6]. In this chapter, an analytical framework for the TFET is developed that reveals a significant short coming of the basic design. A new TFET is proposed that uses dopant engineering and ultra shallow “pockets” of charge to achieve very steep sub threshold slope or swing over many decades of current. This new design permits supply voltage scaling down to V_{dd} of 200 mV when used with appropriate band gap material and is therefore named the “Green” TFET or gTFET. [3.7-3.9]

3.2 A Tunnel Field Effect Transistor (TFET)

It is not entirely obvious how one can create a device that “turns on” by the tunneling process. Figure 3.1 represents the most basic and widely researched embodiment of a transistor that is theoretically not subject to the 60 mV/dec limitation. Whereas in MOSFET, carriers are generated by injection over a gate controlled potential barrier, in this transistor carriers are generated by tunneling through a barrier. In this case, the “barrier” is the semiconductor band gap and the tunneling process is entirely band-to-band, which has been described in detail in Chapter 2. A common name for this transistor is the TFET (Tunnel Field Effect Transistor). Recently, there has been significant research effort by various groups devoted to the TFET as a possible sub 60 mV/dec. low voltage transistor. However, it is important to recognize that the structure shown in Figure 3.1. is not new or novel by itself. Researchers have explored the same structure as a novel device more than a decade ago. [3.1] This leaves much room for improvement with the basic TFET structure, which thus far has not shown significant promise for low voltage operation [3.1-3.6]. Before discussing new designs, the fundamental principle of operation of the TFET must be studied. A simple analytical framework first needs to be developed that agrees well with experimental data. This gives confidence in our model, which can then be used to design a better low voltage, or “greener” transistor.

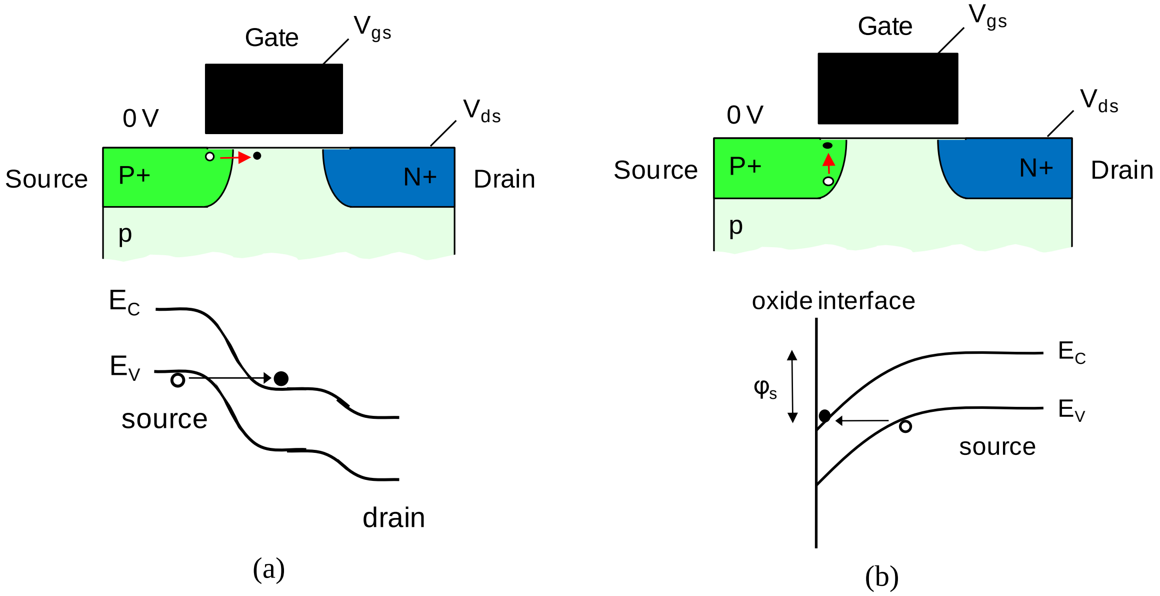


Figure 3.1: One embodiment (the simplest) of a tunnel field effect transistor (TFET). The gate voltage controls the tunnel barrier width, which regulates the amount of current flow in channel. (a) Principle of operation accepted by nearly all researchers where tunneling direction is lateral or parallel to gate dielectric. (b) Principle of operation proposed in this work where tunnel direction is vertical or normal to gate dielectric.

3.3 TFET Analytical Framework

3.3.1 Principle of Operation

Figure 3.1 shows an n-channel TFET structure and energy band diagrams of operation. Figure 3.1(a) shows the principle of operation that almost all researchers have accepted up till this point.[3.1-3.6] A positive gate voltage pulls the energy band downwards in the channel. This allows band-to-band tunneling to occur laterally (parallel to the gate dielectric interface) from source to channel. While this provides at first glance a simple and satisfying description of TFET operation, it is not necessarily the correct picture. Figure 3.1(b) shows another point of view where tunneling is vertical. Gate voltage pulls the energy bands downwards in the gate-source overlap region causing tunneling to occur within the source directed at the gate dielectric. This description of operation is not as intuitive as Figure 3.1(a), but it is one which has been applied to model tunneling leakage currents in MOSFETs very successfully. [3.10] Gate induced drain leakage or GIDL occurs in the drain overlap region of the MOSFET when biased in the “off-state”. For example, p-channel MOSFET GIDL condition with P+ source/drain with gate at V_{dd} and drain at 0 V has the same bias and structure as an n-channel “on state” TFET as shown in Figure 3.2. In addition, the direction of tunneling is one with largest electric field. For an MOS structure the vertical field almost always dominates. Given these arguments, it should not be too surprising that the vertical tunneling viewpoint applies very well to the TFET.

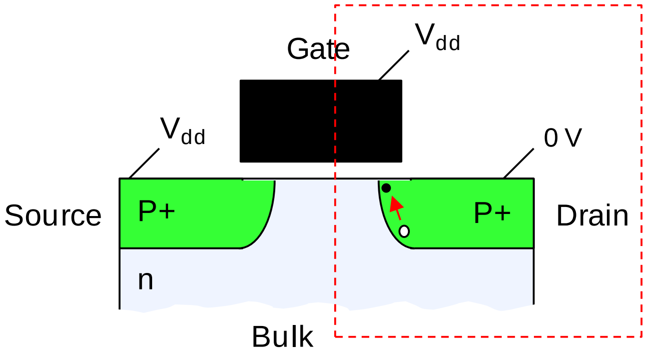


Figure 3.2: A pMOSFET transistor biased in off state regime where gate induced band-to-band tunneling [3.10] occurs is identical in structure and bias to an n-channel TFET. The drain is the TFET source and bulk the TFET drain.

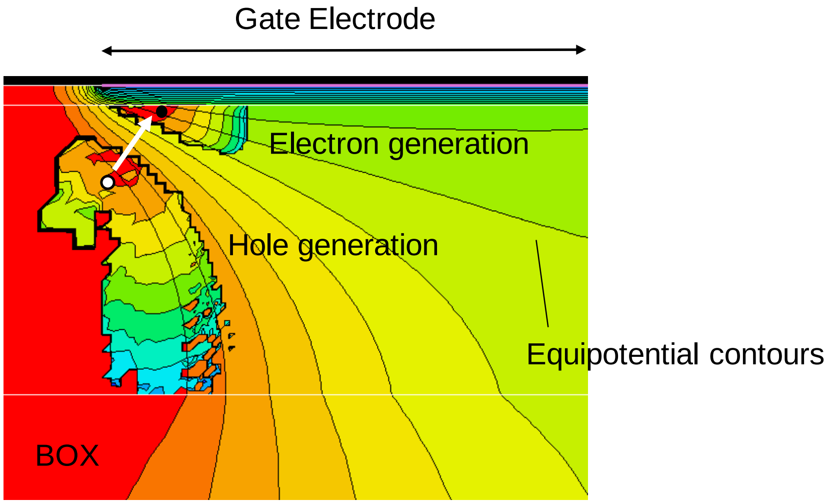


Figure 3.3: Simulation output of the TFET in the on state showing that tunneling direction has both vertical and lateral characteristics and is occurring within the source.

For further confirmation of tunneling direction, Figure 3.3 shows a simulated n-channel TFET in the "onstate". The details of simulator will be discussed later in this chapter. The electrostatic contours are as indicated. The electron and hole generation rate dictates the tunneling direction, which is not entirely vertical or lateral. However, as will be shown in the sections to follow the vertical model agrees very well with experimental data.

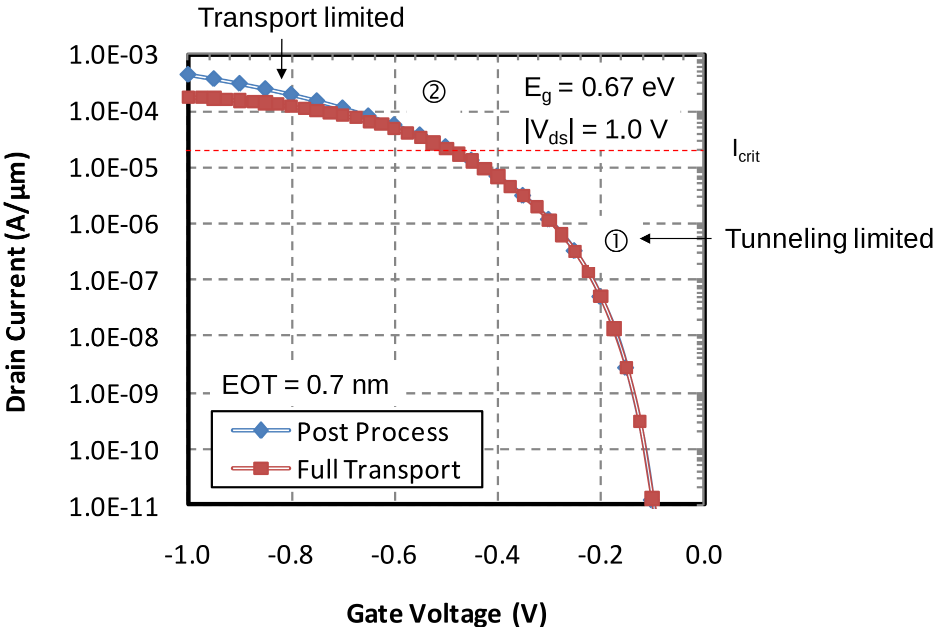


Figure 3.4: I_d-V_g simulation of a typical p-channel TFET shown with E_g=0.67 eV for illustrative purposes. The "Post Process" current results from the solution of Poisson equation only (i.e., no transport). Below approximately 10 µA/µm, the TFET current is entirely tunnel rate limited.

3.3.2 Tunneling Limited Currents

An important TFET concept is demonstrated in Figure 3.4, which is an I_d-V_g simulation of a generic TFET (identical to Figure 3.1) with germanium band gap. The MEDICI device simulation tool with band-to-band tunneling models enabled was used. The exact details of this simulation or simulator will be discussed in later sections. Two IV curves can be recognized. In one case, a regular DC simulation is performed and is labeled as "Full Transport". In this case, the semiconductor device equations are solved simultaneously. Poisson equation is solved self consistently with the drift diffusion transport equations to arrive at the terminal currents. The other curve labeled as "Post Process" is the result of only a solution of the Poisson equation. No transport equations are solved. The current is determined by calculating the tunneling generation rate from the electrostatic potential solution in a post process manner. By integrating the rate (units of $cm^{-3}s^{-1}$) over the tunneling regions and multiplying by charge, units of current are obtained. It should be noted that these two curves are identical below a certain current value. This exact value will depend on the geometry of the device and transport parameters (i.e. mobility) to some extent, however is approximately 10 µA/µm across nearly all TFET simulations. Above this critical current level the curves deviate, with the actual "Full Transport" current falling short of the "Post Process" value. The "Post Process" value is the theoretical maximum current the device can extract assuming infinite carrier mobility. These results suggest there are two regions of operation for a TFET: (1) tunneling limited (2) transport limited. In region (1) the current can be modeled solely by proper account of electrostatics and use of tunneling model. In region (2) the situation is more complex. The presence of large currents alters the electrostatics and therefore the tunneling rate. A self consistency loop must be established.

For the purposes of developing an analytical model, these results greatly simplify the picture. Experimentally all TFET devices have thus far fallen into the tunneling limited regime for useful voltage ranges. Also, the current range where sub 60 mV/dec behavior is of interest is below 10 µA/µm. This suggests that only region (1) needs to be examined in detail for now. For calculation of tunneling limited currents, a correct electrostatic model first needs to be developed.

3.3.3 TFET Surface Potential Electrostatic Model

To model the tunneling limited current correctly the electrostatics must be examined carefully. The vertical tunneling viewpoint is used as the motivation behind the model. This means that tunneling is occurring in the source overlap region in a direction normal to the gate dielectric. The amount of vertical band bending in the source needs to be calculated as a function of the terminal voltages. Once band bending is known the simple tunneling model derived in Chapter 2 can then be used to obtain current as a function of terminal voltages. Figure 3.5(a) shows the setup of the problem. Calculation of band bending is a well known and solved problem in MOS capacitors. For the TFET, however, the situation is a bit different. The source region is in non-equilibrium since under the bias condition outlined in Figure 3.5(a), the source to drain diode is under reverse bias. The electron and hole Fermi levels E_{fn} and E_{fp} will not coincide in the source region. Instead, as shown in Figure 3.5(b) they will be separated by an amount equal to the reverse bias or drain voltage V_{ds}. This is similar to the situation in the depletion region of a simple diode. The additional complexity is because of the addition of the gate terminal. In this n-channel example, E_{fn} will be pushed down relative to E_{fp}, thereby changing the condition of inversion in the source. The effects of both the drain and gate voltage need to be included in the surface potential model. Eq. (3.1) is the well known MOS charge sheet formulation modified for use in the source overlap region and under conditions of non-equilibrium [3.11].

$$V_{gs} = V_{fb,source} + \phi_s + \gamma_{source}\sqrt{\phi_s + \left(kT/q\right)e^{\frac{\phi_s - (2\phi_{f,source} + V_{ds})}{kT/q}}} \tag{3.1}$$

In this equation the last term on the right hand side is the total charge in the semiconductor. The second term under the square root represents the inversion charge. In this case, the condition for inversion is changed from $2\varphi_{f,source}$ to $2\varphi_{f,source}+\mathrm{V}_{\mathrm{ds}}$. This equation must be solved numerically to obtain the surface potential $\varphi_s\left(V_{gs},V_{ds}\right)$ although some approximate closed form expressions are possible. [3.12] Note for large drain bias, the inversion charge term is negligible; meaning all charge on gate is balanced by depletion charge from ionized dopants only. This is equivalent to the condition of “deep depletion” in MOS capacitors.

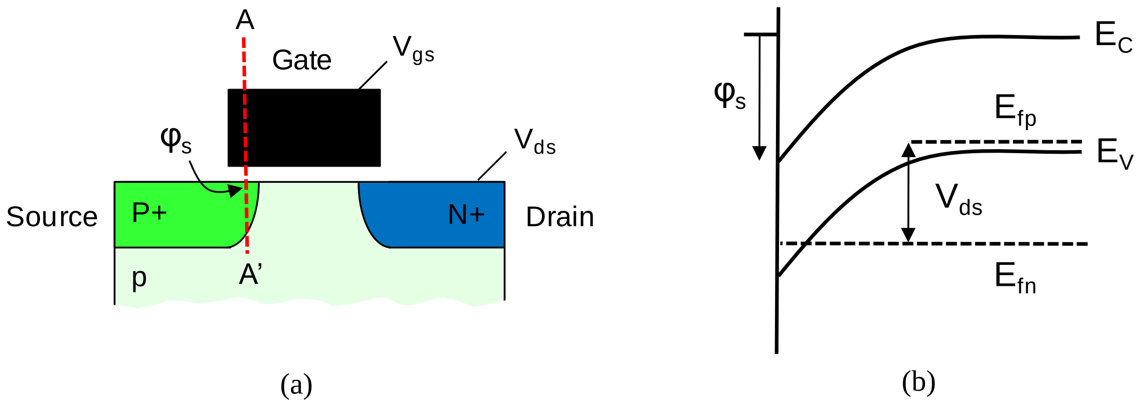


Figure 3.5: (a) Determination of the band bending in the source overlap region allows for calculation of tunneling current. (b) Band diagram normal to the surface showing the separation of Fermi levels resulting from finite drain voltage bias in the tunneling limited regime.

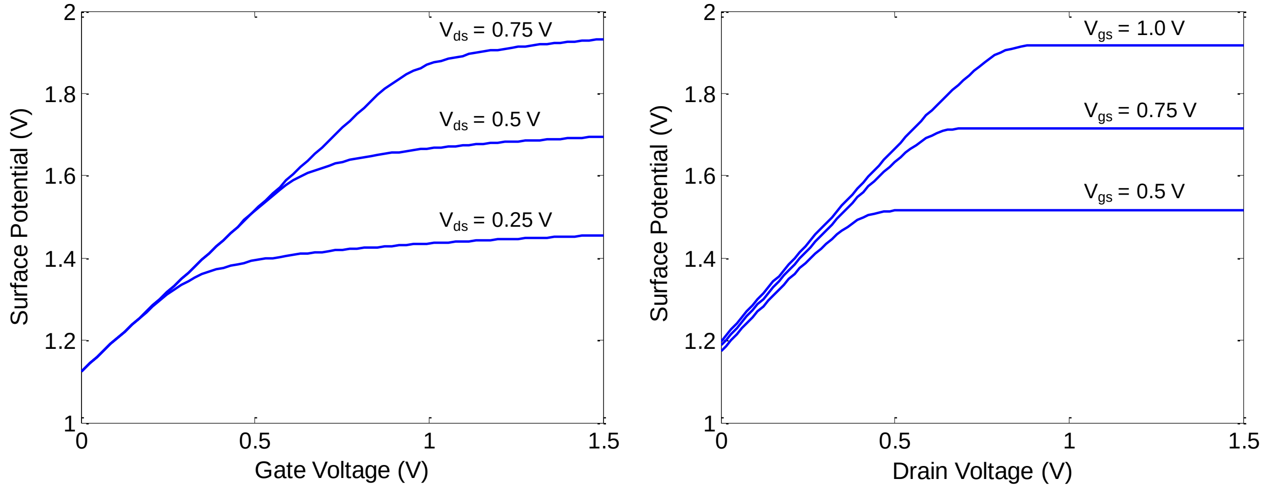
In actuality, the assumption that $E_{fn} = V_{ds}$ in the source region is an approximation valid only in the tunneling limited regime. While $E_{fn} = V_{ds}$ in the drain, the gradient of E_{fn} in channel direction is related to the amount of electron current flow.

$$I_n = qWQ_n\mu_n \frac{dE_{fn}}{dy} \tag{3.2}$$

For smaller electron current (i.e. below the transport limit of Figure 3.4) this gradient is negligible across the channel from drain to source, justifying this assumption of $E_{fn} = V_{ds}$.

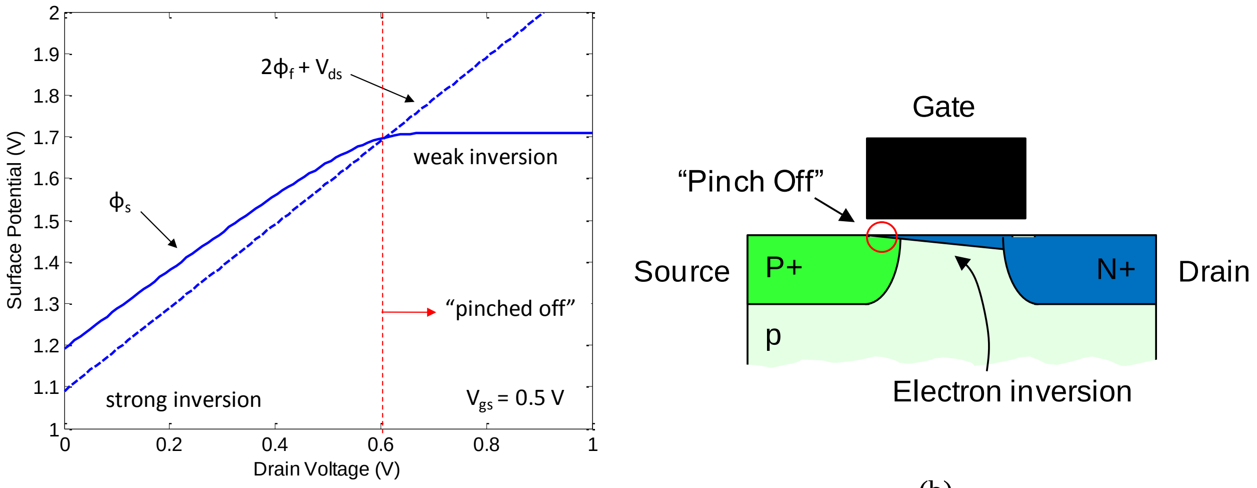
The numerical solutions of Eq. (3.1) in the source overlap are plotted as functions of both drain and gate voltage in Figure 3.6. When the gate voltage is large and drain voltage is small, the relationship between surface potential and gate is almost linear. This is to be expected because in this bias regime, the inversion charge is negligible and the situation reduces a series connection of oxide and depletion capacitance. When gate voltage substantially exceeds the drain voltage, inversion charge will be present which screens the surface potential from further bending. This situation is exactly similar to that of the MOSFET biased into strong inversion.

In Figure 3.6(b) the drain voltage is swept while the gate is held constant. In this case for small V_{ds}, an inversion layer is still present across the channel into the source overlap region. There is a one to one relationship between the drain voltage and surface potential in the source region in this case. Above certain a voltage, the source region inversion charge becomes negligible causing the surface potential to become independent of further drain influence. The source region is “pinched off” causing surface potential saturation in a manner analogous to the MOSFET. Figure 3.7(a) demonstrates this concept quantitatively. From solution of Eq. (3.1), the drain voltage increases the condition for inversion as well as the surface potential up until the pinch off voltage, where the source region begins entering weak inversion causing the potential to stay constant. An analogous viewpoint is that as inversion charge begins to decrease the depletion charge must increase to balance the fixed charge on the gate. This results in an increase of the depletion width or surface potential until inversion charge is negligible.



(a) (b)

Figure 3.6: (a) Numerical solution of Eq. (3.1) as function of gate voltage. For large V_{ds}, the potential is linear with gate voltage similar to MOSFET subthreshold region. (b) Solution of Eq. (3.1) for drain voltage sweep. Above a certain V_{ds}, the potential in source is constant.



(a) (b)

Figure 3.7: (a) From solution of Eq. (3.1), as drain voltage is increased both surface potential and the condition for inversion increase. Beyond a certain drain voltage the source region enters weak inversion and is "pinched off", resulting in saturation of potential. (b) Unlike MOSFET, for TFET "pinch off" occurs on the source side.

The pinch off voltage to first order is approximately $V_{pinchoff} = \left(V_g - V_{t,source}\right) / m_{source}$ as in MOSFET except the threshold and body effect parameter are determined from the source doping concentration instead of body. As shown in Figure 3.7(b), for TFET pinch off occurs in the source rather than near the drain as in MOSFET.

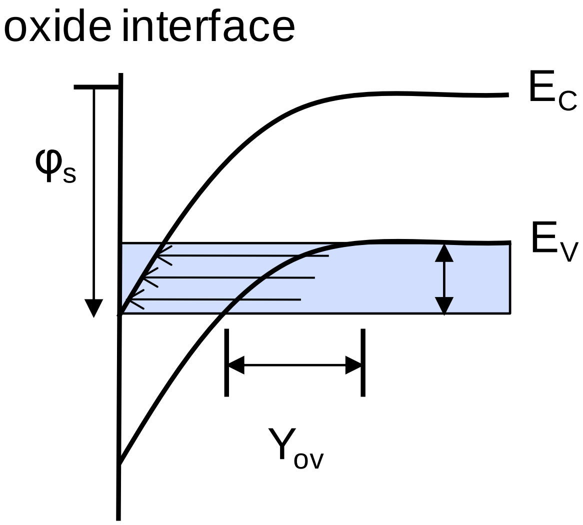


Figure 3.8: Across the overlap (shaded blue), each tunneling path "sees" a different average electrical field.

3.3.4 TFET Analytical IV Model

In the previous section the electrostatics have been modeled as a function of the terminal voltages. More specifically the surface potential in the source is known as a function of gate and drain voltages. Once the surface potential or vertical band bending is known the problem now amounts to that shown in Figure 3.8. For given φ_s and doping concentration in source the bend bending profile can be calculated. The proper tunneling model derived from Chapter 2 needs to be applied to the situation. The local tunneling model with average electric field across tunneling path is shown in Eq. (3.3)

$$G_{BTBT} = AE_{avg}^2 \exp\left(-\frac{B}{E_{avg}}\right) \tag{3.3}$$

To obtain units of current, Eq. (3.3) must be integrated over all possible tunneling paths shown as the shaded blue region in Figure 3.8. However, the average tunneling electric field varies across all the possible paths. The generation rate must be integrated, however, to obtain a closed form expression an approximation can be made.

$$I_{ds}(V_{gs},V_{ds}) = qWX_{eff} \int_{o}^{y_{ov}(\phi_s)} G_{BTBT}(E_{avg})dy \approx qWX_{eff} Y_{ov} G_{BTBT}(E_{avg,ov}) \tag{3.4}$$

An average of the average tunneling field across all paths can be calculated from simple electrostatics for this situation of uniform doping concentration. The tunneling rate is then assumed constant over the blue region with this value of electric field. The top integration limit is the amount of overlap in position, which can also be calculated from MOS theory with knowledge of the surface potential.

$$E_{avg,ov} = \frac{2}{3}\sqrt{\frac{qN_{source}}{2\varepsilon_{si}}}\,\frac{1}{(\phi_s - E_g)}\left[\phi_s^{3/2} - E_g^{3/2} + \left(\phi_s - E_g\right)^{3/2}\right] \tag{3.5}$$

$$Y_{ov} = \sqrt{\frac{2\varepsilon_{si}\left(\phi_s - E_g\right)}{qN_{source}}} \tag{3.6}$$

This allows the integration to be performed in closed form. The final result is shown in Eq. (3.7).

$$I_{ds} = \frac{4}{9}q\cdot X_{eff}\cdot\sqrt{\frac{qN_{source}}{2\varepsilon_{Si}}}\cdot\frac{1}{\left(\phi_s - E_g\right)^{3/2}}\cdot\left(\phi_s^{3/2} - E_g^{3/2} + \left(\phi_s - E_g\right)^{3/2}\right)^2 A\exp\left(-B\frac{3}{2}\cdot\sqrt{\frac{2\varepsilon_{Si}}{qN_{source}}}\cdot\frac{\phi_s - E_g}{\phi_s^{3/2} - E_g^{3/2} + \left(\phi_s - E_g\right)^{3/2}}\right) \tag{3.7}$$

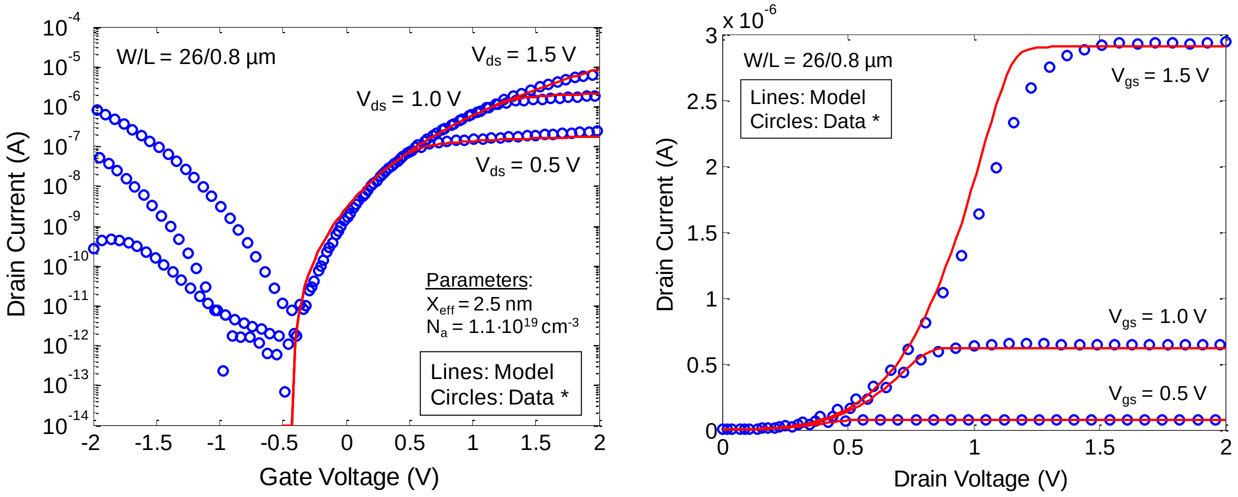
X_{eff} is the length of the source overlap region that is contributing to the tunneling current. This is a physics based closed form analytical surface potential model of the TFET current when in the tunneling limited region (i.e., currents below approximately 10 µA/µm). This model is continuous and valid from linear (below pinch off) to saturation regime in the output characteristics. Note that Eq. (3.7) is valid only when band bending is in excess of the band gap (tunneling overlap condition). When there is no overlap the drain current is equal to the floor leakage of the device (reverse bias diode leakage). X_{eff} and N_{source} are treated as fitting parameters in this model as there is no way to determine this from an experimental device. The tunneling parameters *A* and *B* are set to calibrated values determined from literature. [3.13]

3.3.5 Verification of the TFET Analytical Model

The model in Eq. (3.7) is compared to experimentally measured TFET devices both fabricated within Microlab and those of other researchers reported in literature. The *A, B* parameters of the tunneling model are defined as follows, which are the calibrated for silicon values reported in literature and used as default in device simulators. [3.13]

$$\begin{aligned} A &= 3.5\cdot 10^{21}\ \mathrm{cm^{-1}s^{-1}V^{-2}} \\ B &= 22.5\cdot 10^{6}\ \mathrm{V/cm} \end{aligned} \tag{3.8}$$

Figure 3.9 are measurement results from a fabricated n-channel TFET by colleague Anupama Bowonder that is structurally identical to Figure 3.1. A 3 nm grown SiO_2 layer with poly silicon gate was used as the gate stack. The P+ and N+ source and drain were defined by ion implantation aligned to the gate edge. Figure 3.9 shows very good agreement of the analytical model with the experimental data. The fitting parameter values N_{source} and X_{eff} are indicated in the figure. The output characteristics also agree very well with the model.



(a) (b)

Figure 3.9: (a) Comparison of the model with experimental TFET data from Berkeley. [3.14] I_d-V_g curve shows excellent fit. (b) The corresponding I_d-V_d also agrees well with the model derived in this section. X_{eff} and N_{source} are treated as fitting parameters.

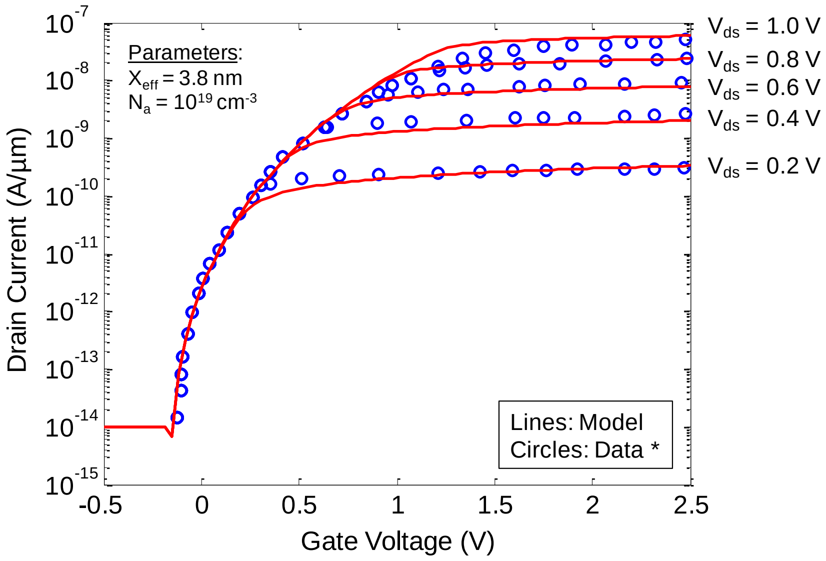


Figure 3.10: Comparison of model with experimental data from other researchers also shows good fit. [3.15]

Note that the non-linear rise in current for low drain voltage is a characteristic of the TFET that the model captures. This arises because for low drain bias, the surface potential in the source is modulated directly by the drain voltage. The relationship between electric field and potential is to the power of 1/2. Tunneling current itself depends exponentially on the field from Eq. (3.3). It should, therefore, not come as surprise that the I_d-V_d behavior is very non-linear at low the V_{ds} regime.

The analytical model also shows good fit to reported TFET data from other researchers as shown in Figure 3.10. Note that drain current saturation on the I_d-V_g curve is captured with the model. This saturation effect arises when the transistor becomes biased below pinch off or when an inversion layer begins to form in the source during the sweep. The formation of this layer reduces the coupling of the gate voltage to surface potential by screening effects. Biasing the device with even larger V_{ds} ensures the source remains "pinched off" during the gate voltage sweep. These effects typically arise when the pinch off voltage is very large. In this device the threshold is negative resulting in a pinch off voltage that is always larger than the applied gate voltage.

The good agreement of the analytical model with both in house and other reported data gives strong confidence in its overall correctness. These models have centered on the vertical tunneling viewpoint for TFET operation, which has now been shown to be accurate. The model can now be used to identify the short comings of the generic TFET structure in Figure. 3.1. The outcome of this exercise will be a superior transistor design, i.e. one that achieves larger current with smaller supply voltage or simply stated one that is "greener".

3.3.6 Transport Limited Models

The effects of transport become important for drain currents above approximately 10 µA/µm. However, modeling these effects is not trivial. The simplest method was described in section 3.3.1 using the "post process" method with a device simulation tool. A general TFET structure in Figure 3.1 is simulated with full transport models. A second curve is generated by simulating again without transport by assuming infinite mobility and no velocity saturation. This second curve will begin to deviate from the first at approximately 10 µA/µm. The ratio of the current values of these two curves is the transport degradation factor showing in Figure 3.11. The exact shape of this curve will depend to some extent on geometry and mobility parameters. This factor $f_{transport}$, which will be dependent on current, allows for a first order transport correction to the tunneling limited model in Eq. (3.7).

$$I_{ds,transport} = f_{transport}\left(I_{ds}\right)\times I_{ds}\left(V_{gs},V_{ds}\right) \text{ of Eq. (3.7)} \tag{3.9}$$

Developing a physics based transport model, however, is much more complicated. The situation in the source region requires a self consistency loop. For a given current flowing through this region, some amount of charge will be present. This charge alters the electrostatics such that the total band bending and electric field is lower, thereby decreasing the tunneling current flowing through the source. This concept is illustrated in Figure 3.12.

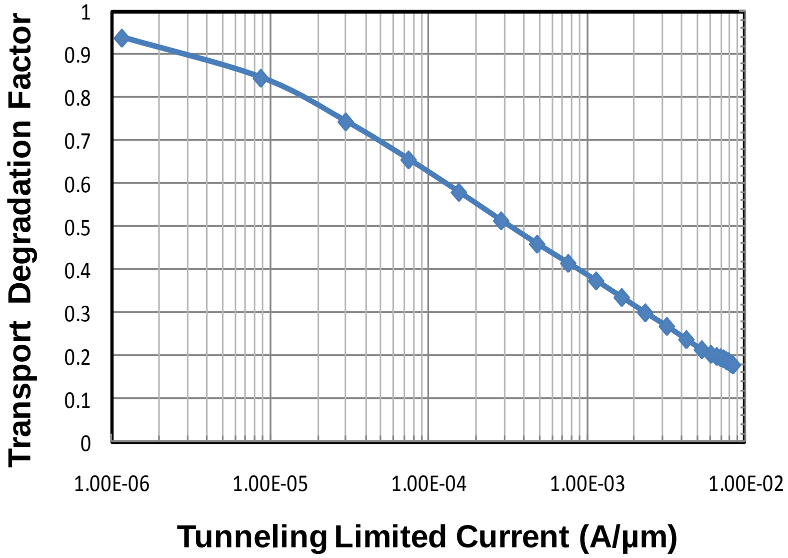


Figure 3.11: A simple transport model for TFET uses concept of degradation factor. This is the ratio of full transport to post process current for a typical TFET. The exact shape of this curve depends to some extent on device geometry and mobility parameters.

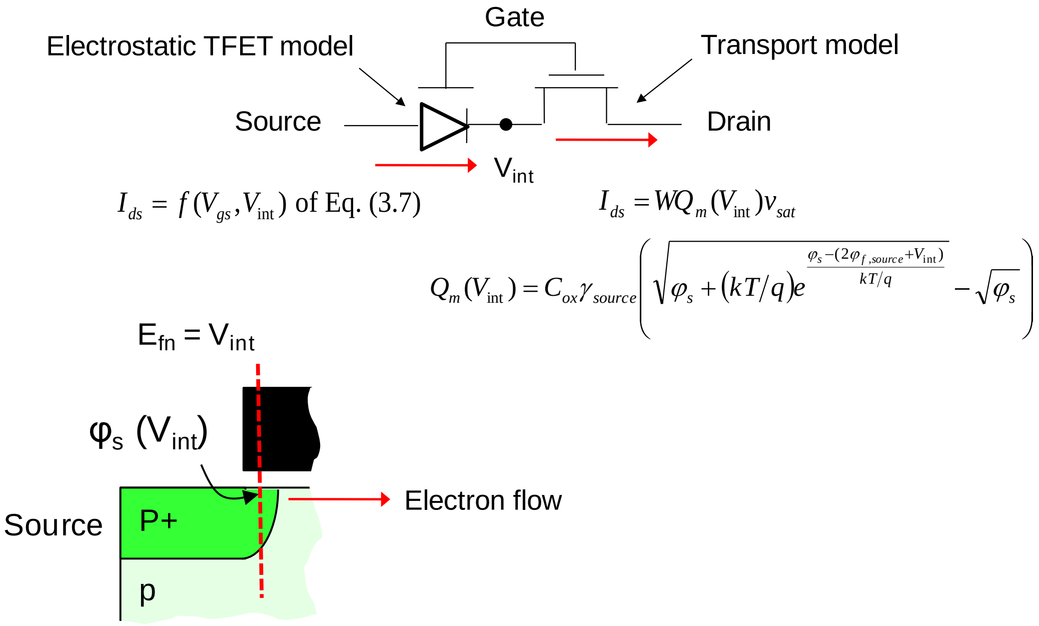
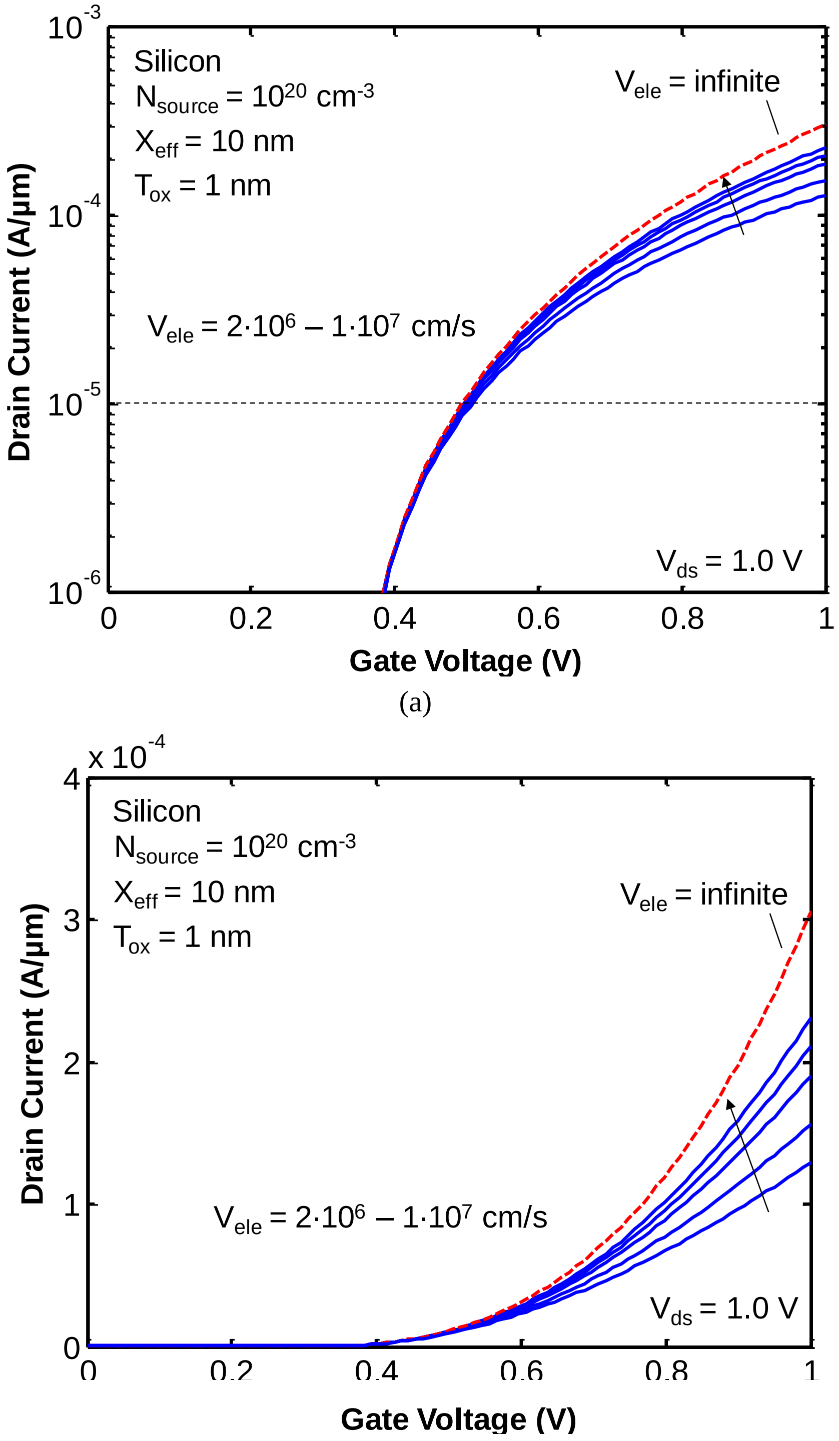

Gate
Electrostatic TFET model
Transport model
Source
Drain
V_{int}
$I_{ds} = f(V_{gs}, V_{\mathrm{int}})$ of Eq. (3.7)
$I_{ds} = WQ_m(V_{\mathrm{int}})v_{sat}$
$Q_m(V_{\mathrm{int}}) = C_{ox}\gamma_{source}\left(\sqrt{\varphi_s + (kT/q)e^{\frac{\varphi_s - (2\varphi_{f,source} + V_{\mathrm{int}})}{kT/q}}} - \sqrt{\varphi_s}\right)$
$E_{fn} = V_{int}$
φ_s (V_{int})
Source
P+
p
Electron flow

Figure 3.12: A transport model for the TFET uses two elements in series. The first element is the electrostatic TFET model of before. The second is a parasitic MOSFET. The internal voltage is the location of the electron Fermi level in the source and must be determined self consistently from Kirchhoff's current law.

This is a two element model, where the second element is the transport model depicted in the figure as a series MOSFET. The internal node V_{int} must be determined self consistently such that the tunneling current of Eq. (3.7) and the transport current are equal. V_{int} replaces V_{ds} in Eq (3.7) and determines the amount of charge Q_m in the transport model. The carrier velocity is assumed to be the saturation velocity v_{sat} to first order although can be treated as a parameter. For ultra short channel length devices (L_{gate} on the order of the electron mean free path) it is possible the electron velocity may exceed the saturation value. This is desirable for largest drive current, since higher velocity implies less Q_m for same value of current. The limit of zero Q_m is the situation of maximum possible current since this corresponds to $V_{int} = V_{ds}$. This TFET transport model with self consistency loop has been implemented in MATLAB. Figure 3.13 shows the model I_d-V_g of an optimized TFET with various electron velocity values. The case of infinite velocity corresponds to the “Post Process” or tunneling limited regime as discussed in section 3.3.2. Lower electron velocity results in larger degradation of current compared to the tunneling limited value as expected. It can also be seen that transport effects begin to appear at drain current of approximately 10 µA/µm for reasonable values of electron velocity. This is in close agreement with the actual TFET simulation results in 3.3.2. It should be noted that achieving convergence in this self consistent loop is challenging for larger values of current or smaller values of electron velocity.



(a)

(b)

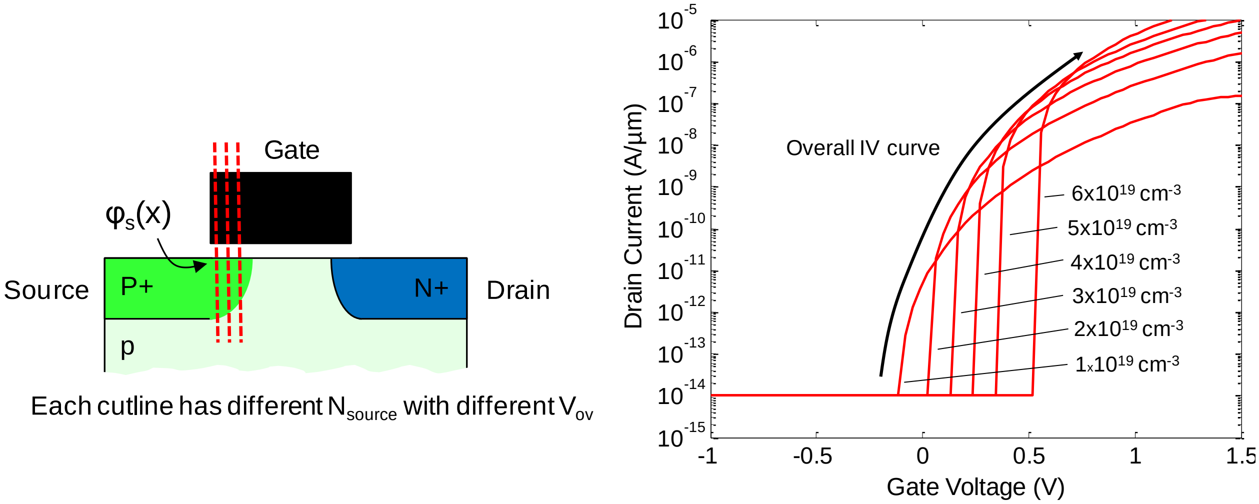
Figure 3.13: Physics based TFET transport model I_d-V_g with various electron velocity values. The case of infinite velocity corresponds to the tunneling limited current discussed in section 3.3.2. (a) Log scale (b) Linear scale.

3.4 Limitations of the Simple TFET

So far there has been no mention of the lateral doping gradient of the source region of the general TFET in Figure 3.1. The models developed thus far have assumed uniform doping concentration within the source and perfect termination at the edge. The tunneling limited model only has one doping N_{source} as a parameter. In actuality, the doping profile is never perfectly abrupt in the lateral direction and always has some finite gradient. In Figure 3.14(a) this lateral gradient is modeled as various tunneling segments in parallel indicated by the vertical cut lines. Each cut line samples a different N_{source} because of the lateral gradient. The net current response will be a summation of each contributing segment with each having a different overlap voltage V_{ov}. The overlap voltage is a terminology that will be used consistently throughout the rest of this section and remaining chapters. Simply stated it is defined as the gate voltage at which there is sufficient band bending for band-to-band tunneling to occur. As shown in Figure 3.14(b), the lighter doping concentration segments will "turn on" because they have a lower overlap voltage. This presents a challenge to the analytical model since as mentioned above only one N_{source} is specified. The simplest modification is to allow the doping concentration N_{source} to become bias dependant. One implementation is to treat N_{source} as a linear function of the band bending in excess of the band gap. Conceptually this makes sense because at higher bias or electric fields, the heaver doping segments contribute the most current.

$$N_{a,eff}\left(\varphi_s\right) = N_{low} + \left(N_{high} - N_{low}\right)\frac{\left(\phi_s - E_g\right)}{\phi_{s,high} - E_g} = N_{low} + \eta\left(\phi_s - E_g\right) \tag{3.10}$$

Figure 3.15 shows the improved fitting to a simulated generic n-channel TFET with significant source doping gradient. A single value of N_{source} is not able to fit the simulated results.



(a) (b)

Figure 3.14: (a) The graded region of the TFET source may be treated as independent segments each with different doping and turn on or overlap voltage V_{ov}. (b) The I_d-V_g curve is the sum of all segments, which results in "gradual" turn on from the lighter doped segments.

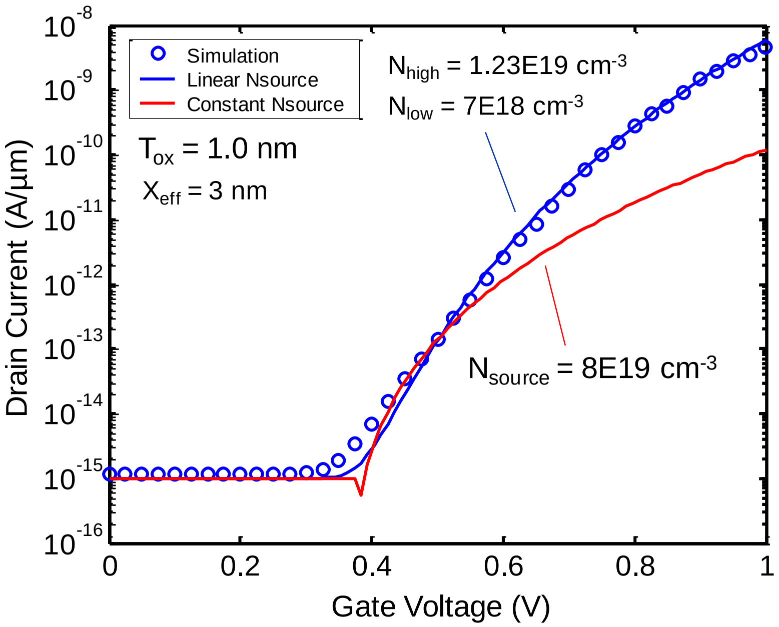
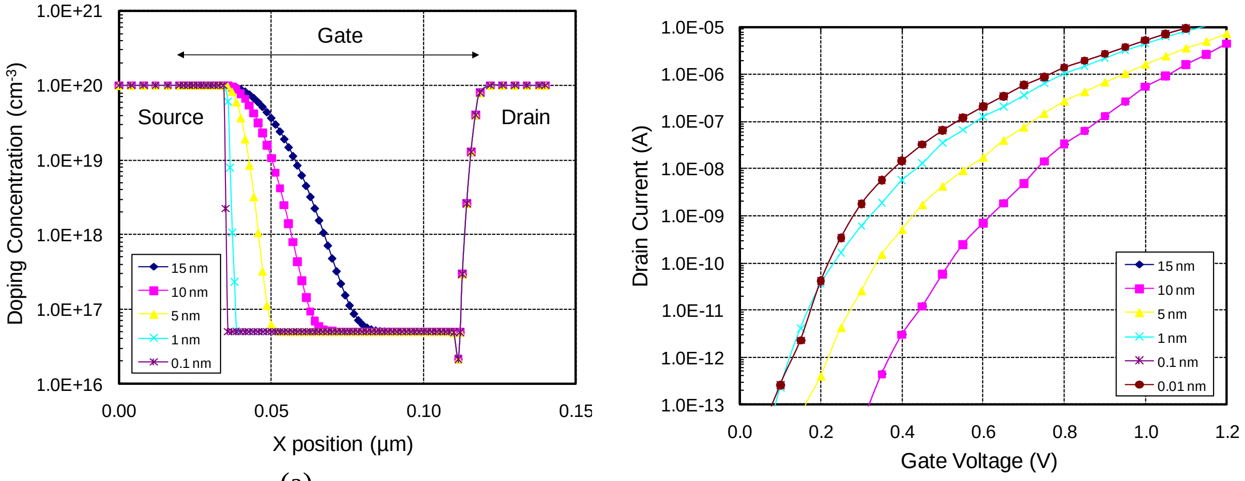


Figure 3.15: The graded doping effect can be taken into account in the model by making the N_{source} a linear function of surface potential. This modification fits a graded source TFET simulation very well. Note that the single N_{source} model is unable to agree with the simulation results.

Most troubling from this discussion is that the lateral source doping gradient ultimately results in a “gradual” transistor turn on characteristics. Tunneling initially starts occurring in the lighter doped segments near source edge (where V_{ov} is lowest) where the electric field is also small. From Figure 3.14(b) the lighter doped segments result in the least desirable turn on characteristic. Since the overall IV curve is the summed response of all segments, the I_d-V_g will not be steep over many decades. Even if the generic TFET transistor can be engineered somehow to achieve sub 60 mV/dec swing, it will only occur over a very small range of current. This observation has been seen in various reported TFET experimental measurements. [3.15]

Figure 3.14(b) reveals important information on how a better transistor can be designed. The IV curves of the lighter doped source segments are undesirable. However, the I_d-V_g of the 5E19 cm^{-3} and above case is very desirable. A very steep swing is seen over many decades of current. For these heaver doped segments, because the condition for overlap or V_{ov} is larger, the electric field at the overlap condition is also larger. This results in a sudden and rapid increase in current as tunneling is permitted from the overlap of the conduction and valance band of an already “thin” tunnel barrier. In the lighter doped segments, the electric field is not very large at V_{ov} resulting in very little jump in current. This steep swing behavior over many decades of current is called the “sudden overlap” effect. This results from the presence of the energy gap, which permits the tunneling process to be completely “turned off” when electrons in the valance band no longer have states to tunnel into on the receiving side. When the band bending is less than the band gap the tunnel current is zero. Once bands are overlapped, the transistor swing is determined by modulation of the tunnel probability with gate voltage, which is seen to be not very steep.



(a)

(b)

Figure 3.16: (a) Simulation of a graded source TFET with various levels of abruptness. (b) Corresponding I_d-V_g curve. Even a perfectly abrupt source does not result in steep turn on that is seen in the model of Figure 3.14 for the heavily doped segments.

To design a better transistor two challenges need to be overcome. (1) The lighter doping segments of Figure 3.14(b) need to be suppressed somehow. (2) The V_{ov} of the heavier doped segment needs to be lowered to useful values. To suppress the lighter doping segment "turn on", making the source profile more abrupt might help. An abrupt lateral profile has less lighter doping segment contribution than a graded one. The actual device simulation results shown in Figure 3.16, however, are intriguing at first.

A general n-channel TFET of Figure 3.1 in silicon is simulated (using MEDICI device simulator) with the local band-to-band tunnel model with average electric field using calibrated A and B parameters as explained in Chapter 2. The gate oxide thickness in this case is 1 nm. Unless otherwise stated full transport is enabled. Various lateral source doping abruptness is simulated. The figure to the right is the corresponding I_d-V_g outputs for each doping profile. As the profile is made more abrupt the swing of the I_d-V_g curve is improved to an extent. However, no "sudden overlap effect" such as that shown in the simple model of Figure 3.14(b) for the 5E19 cm^{-3} segment is observable. Even when the doping profile is perfectly abrupt, the turn on is still mostly "gradual" with a swing of approximately 50 mV/dec occurring only from 10^{-12} to 10^{-11} A/µm.

Clearly, there is some breakdown in the individual segment model of Figure 3.14(b) when the doping profile is made hyper abrupt. This model has assumed one dimensional electrostatics in the direction normal to the gate dielectric. In actuality, a two dimensional Poisson equation needs to be solved. Qualitatively speaking, the transistor effectively "sees" some average or effective doping near the source edge when the doping profile is abrupt as a result of the 2D Poisson equation. Although this explanation is not quantitatively satisfying, the simulation result of Figure 3.16 is very clear. Whether the lateral source termination is graded or abrupt, band-to-band tunneling always first occurs at the source edge. The overlap voltage V_{ov} is always smallest in the source edge region. It is not possible to suppress the lighter doping segments from "turning on" with the basic TFET structure of Figure 3.1. This unwanted characteristic is called "source edge tunneling".

3.5 A "Green" Tunnel Field Effect Transistor (gTFET)

To solve some of the issues of the general TFET discussed in the previous section an alternate design is proposed as shown in Figure 3.17. As will be shown this design permits larger on current at lower supply voltages when compared to the standard TFET. Circuits comprised of the transistor in Figure 3.17 will be of much lower power consumption to that of those comprised entirely of Figure 3.1 when operating at same performance or speed. One could argue that this proposed design is more energy efficient or "greener". Mostly to serve as a distinction compared to the prior work and general TFET structure, this transistor is named the "green" TFET or gTFET for short.

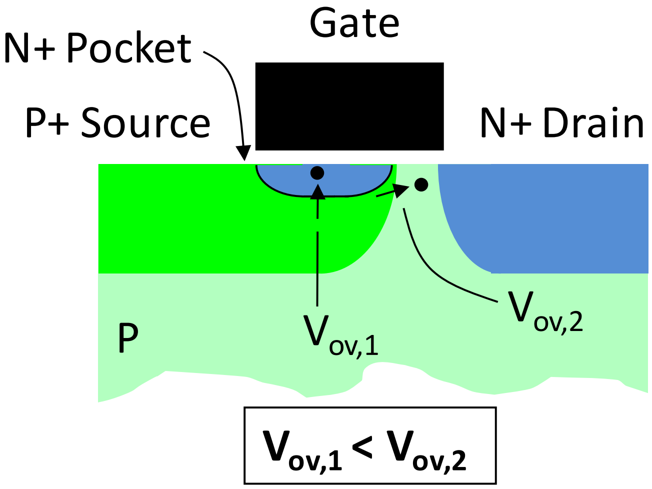


Figure 3.17: Structure of the proposed gTFET. Two possible tunneling paths exist in this device. $V_{ov,1}$ corresponds to tunneling from source to pocket which results in steep swing. $V_{ov,2}$ corresponds to tunneling occurring in the source edge region which results in "gradual' turn on. A fully depleted N+ pocket of charge is designed to ensure $V_{ov,1} < V_{ov,2}$.

As was discussed in the previous section, there is no way to suppress the onset of "source edge tunneling". However, it is possible to completely over shadow the effect. In Figure 3.17 the simple TFET is modified by increasing the amount of source overlap with the gate. An ultra thin fully depleted pocket of charge is introduced atop the overlap region. Two possible tunneling paths exists in this structure: (1) Band-to-band tunneling from uniform and heavily doped source to pocket indicated by overlap voltage $V_{ov,1}$. (2) Tunneling at source edge with $V_{ov,2}$. Ordinarily without the introduction of the N+ pocket, $V_{ov,1}$ would be much larger than $V_{ov,2}$ since the doping concentration in the source region is much larger than the effective doping at the edge. Not surprisingly this would have resulted in a gradual "turn on" that is characteristic to source edge tunneling as discussed in the previous section. However, the fully depleted N+ pocket serves as a thin sheet of positive charge that effectively lowers the flat band voltage in the source overlap region. With correct amount of charge, the flat band voltage can be lowered such $V_{ov,1}$ is less than $V_{ov,2}$. When this condition is satisfied the effects of source edge tunneling are effectively overshadowed by source to pocket tunneling. A steep switching I_d-V_g corresponding to the larger than 5E19 cm^{-3} segments of Figure 3.14(b) is possible. Note that for a p-channel gTFET (not shown) all doping and voltage polarity are simply reversed.

There are three main advantages of the gTFET over the ordinary TFET. (1) When engineered properly (i.e. $V_{ov,1} < V_{ov,2}$) the condition for overlap occurs in the high electric field P+ source region, resulting in rapid rise of current and consequently steep swing over many decades of current. (2) The tunneling area and therefore drive current can be controlled by the length of the pocket. This results in significant on current enhancement since pocket length can be 10X or 20X larger than effective area for tunneling in a TFET. (3) The turn on or overlap voltage V_{ov} of the gTFET can be controlled or adjusted by dose of charge in the pocket.

The simulated energy band diagram is shown in Figure 3.18 in both the “on” and “off state”. During the “off” state there is no overlap, hence zero tunneling current. When the device is turned “on” the gate raises the potential in the N+ pocket through capacitive coupling, causing valance band electrons to tunnel from the P+ source to the pocket. The generated electrons drift to the drain to be collected as drain current.

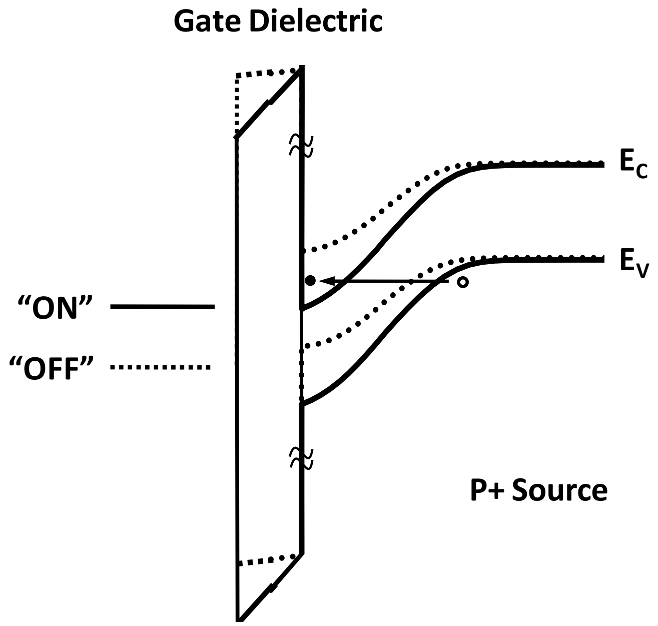


Figure 3.18: Energy band diagram of the gTFET calculated normal to the gate dielectric. In off state there is no overlap between the valance and conduction bands.

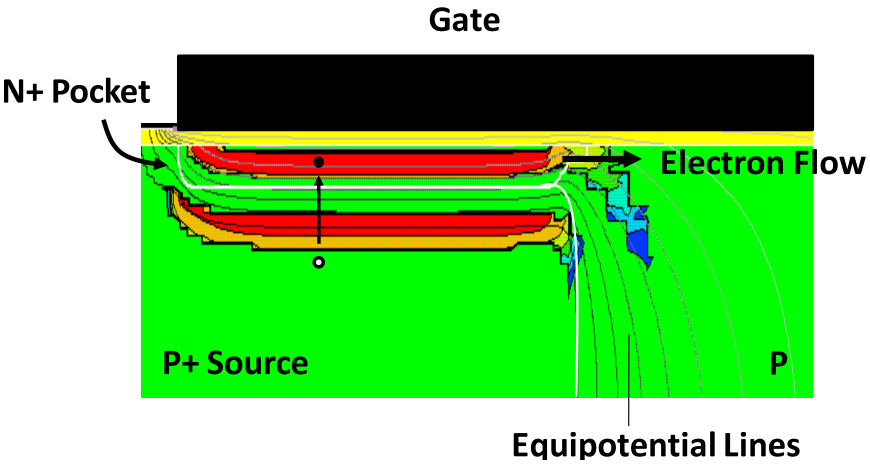
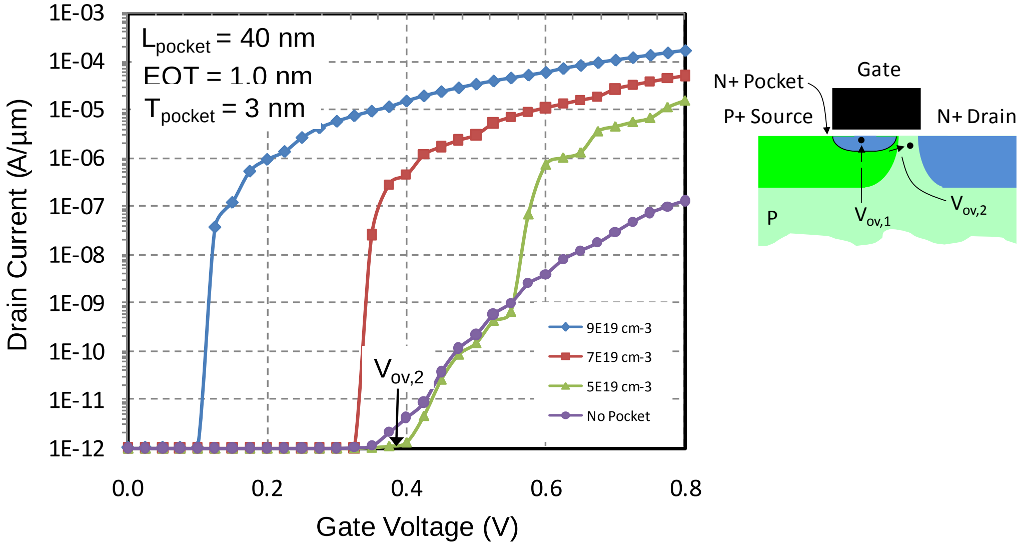


Figure 3.19: Simulation output of the gTFET in the on state. The tunneling generation rate occurs uniformly across the pocket length.

Figure 3.19 is the output of the device simulation showing both the electrostatic potential contours and the tunneling generation rate. Holes and electrons are generated at the start and end of the tunneling path respectively as a result of the band-to-band tunneling process. The generation rate is uniform across the length of the pocket resulting in significant drive current enhancement as will be demonstrated in the following sections.

3.5.1 Simulation of the gTFET Design Space

For the following simulations the MEDICI device simulator is used unless otherwise specified. The local tunneling model of Chapter 2 with average electric field is used. The A and B tunneling parameters are left at the default calibrated values as reported in Eq. (3.8). Unless otherwise specified Poisson-Continuity-Transport equations are self consistently solved to obtain transistor currents. Effects of quantum confinement on tunneling are not included, which is not possible with the current TCAD tools. All simulations are done in silicon unless stated differently. At higher values of drain current obtaining convergence for self consistent solutions can be very difficult. For some situations a “Post Processor” method is used to obtain the IV curve as described in the beginning of this chapter and will be explicitly stated. For the most part n-channel gTFETs are simulated. However, some p-channel simulations are shown when convergence problems with n-channel are too severe. From experience, the p-channel device has always been “easier” to simulate. The source and drain doping concentration are fixed at 1E20 cm^{-3} throughout all simulations unless otherwise specified.



$$\boxed{\Delta V_{ov,1}\ \alpha\ qN_{dose,pocket}/C_{ox}}$$

Figure 3.20: Simulation of gTFET for various different pocket doping concentration. The effect of the pocket is to introduce a flat band voltage shift. When pocket is not doped heavily enough turn on characteristics degrade as shown from the 5E19 cm^{-3} curve.

An n-channel gTFET simulation is shown in Figure 3.20 where the pocket doping concentration is varied. The pocket thickness is kept constant at 3 nm. The amount of flat band shift is related to the dose in the pocket as is seen with decreasing V_{ov} with increasing pocket concentration. The general TFET of Figure 3.1 is shown in the “No Pocket” curve, whose characteristics are significantly worse than those of the gTFET. The “No Pocket” curve is the I_d-V_g of tunneling path 2 with $V_{ov,2}$, i.e., source edge tunneling. For larger pocket concentration this “gradual” turn on is hidden by the “sudden overlap” of source to pocket tunneling, i.e., $V_{ov,1} < V_{ov,2}$. However, when pocket concentration is decreased to 5E19 cm^{-3}, the turn on is identical to the “No Pocket” case until the source to pocket tunneling dominates. This demonstrates that when the pocket dose is not large enough or when $V_{ov,2} < V_{ov,1}$ the swing is severely degraded.

In Figure 3.21, the amount of drive current degradation for constant pocket dose with varying pocket thickness is shown. The current is calculated for fixed overdrive voltage $V_g - V_{ov}$. As the pocket is made thicker, the drive current is degraded because the effective oxide thickness is essentially larger. The location of peak tunneling, which occurs at the pocket to source junction, is pushed further from the gate dielectric interface. It is also important that the pocket remain fully depleted otherwise an undepleted layer of N+ is possible near the surface resulting in a parasitic MOSFET limiting swing to above 60 mV/dec. Above junction depth of 6 nm the simulation fails to converge, suggesting that this is approximately the border between fully and partially depleted pocket. From these simulations it is clear a heavily doped and ultra shallow pocket junction is required for optimal gTFET performance.

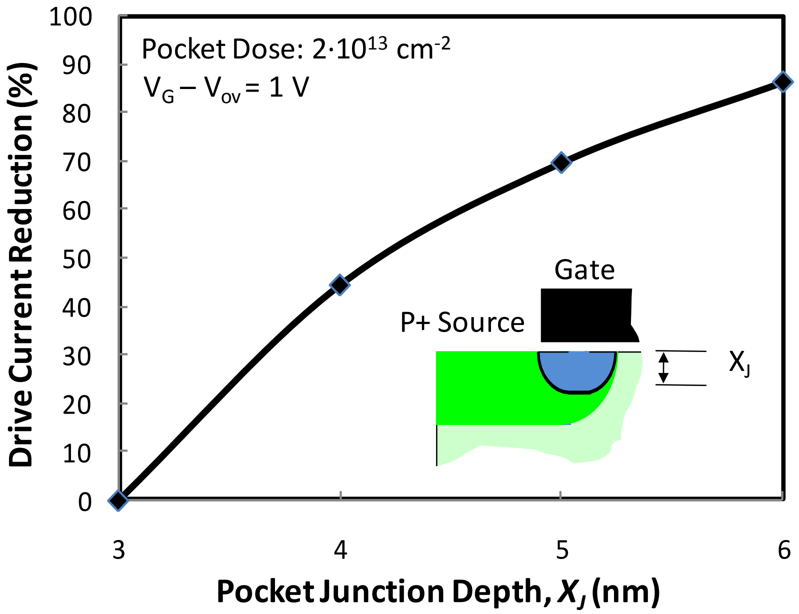


Figure 3.21: Simulation of gTFET for various pocket junction depth for fixed overall pocket dose and overdrive voltage. As the junction is made less shallow, the drive current degrades.

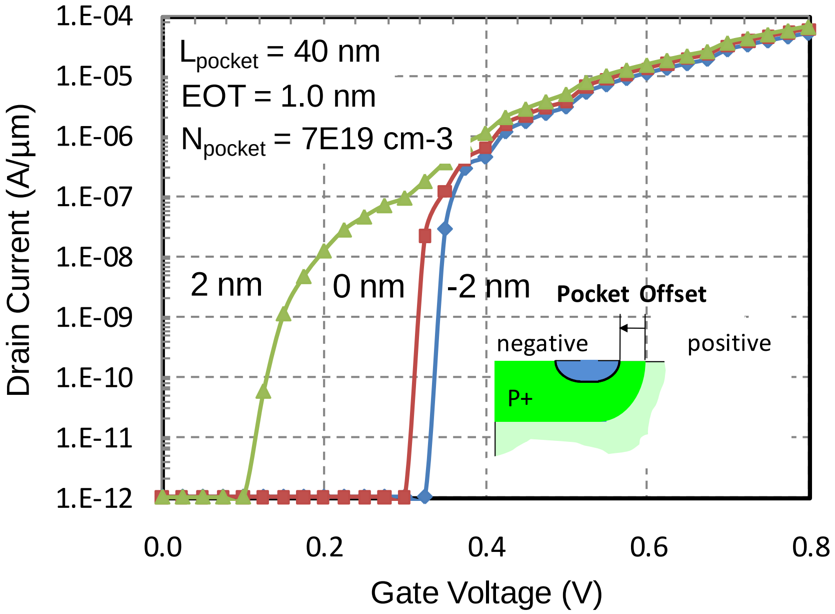
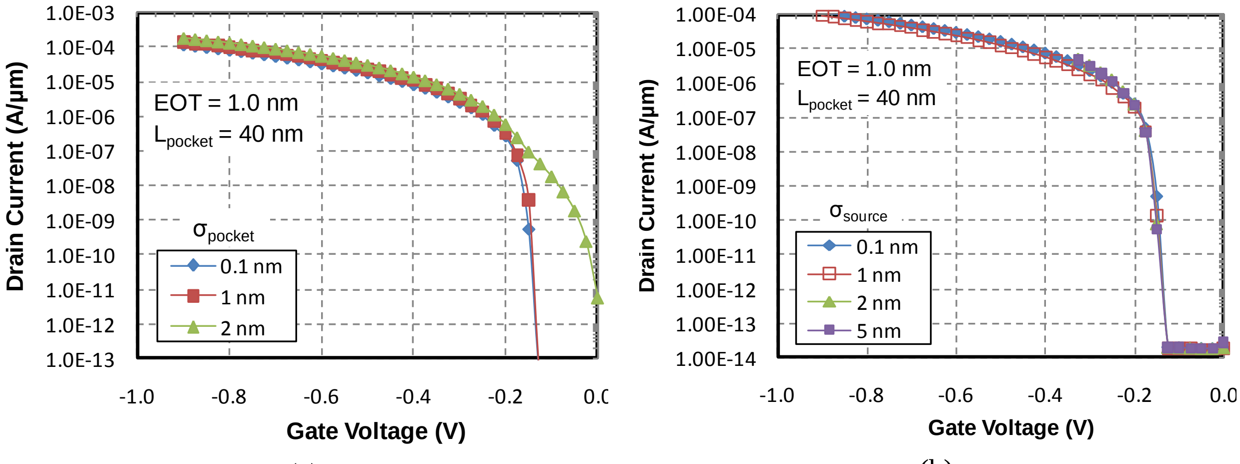


Figure 3.22: Simulation of the pocket offset for gTFET. When pocket extends beyond source edge the swing is degraded.



(a) (b)

Figure 3.23: Effect of gTFET pocket and source lateral abruptness. (a) Source profile is held fixed while pocket profile is varied. A graded pocket profile results in poor turn on characteristics. (b) Impact of source lateral profile is not significant.

The impact of pocket lateral position is shown in Figure 3.22. When the pocket extends beyond the source (positive offset), the I_d-V_g curve is severely degraded. This can be understood qualitatively as follows. Since the role of the pocket is to shift the flat band voltage or lower the V_{ov}, when pocket extends beyond source the overlap voltage $V_{ov,2}$ of source edge tunneling is lowered along with $V_{ov,1}$ by the same amount. This means that it will be impossible to achieve the $V_{ov,1} < V_{ov,2}$ condition for “sudden overlap” swing. When the pocket is instead enclosed by the source (negative offset) the swing is unaffected. Negative offset beyond 2 nm results in non convergence issues and the possibility of forming a barrier to electron flow out of the pocket. The challenges of potential fabrication of the gTFET are brought to immediate attention in these three simulations. A heavily doped, ultra shallow, and near perfectly aligned pocket is required to achieve “sudden overlap” steep swing.

3.5.2 Examination of gTFET Source and Pocket Abruptness

The simulations thus far have assumed a fairly abrupt pocket and source doping Gaussian gradient of $\sigma = 1$ nm. In Figure 3.23(a) the impact of the pocket doping gradient is examined in more detail. In this case the lateral source gradient is held fixed at its nearly abrupt value of $\sigma = 1$ nm, while pocket lateral gradient is varied. A p-channel gTFET is simulated in this case because of convergence challenges; however the trends should be equally valid for the n-channel gTFET. As the doping profile is made less abrupt the I_d-V_g characteristic is severely degraded. However, for Figure 3.23(b), where the source profile is varied while pocket is held constant, the turn on characteristic is mostly independent of the source lateral abruptness. In this case, as long as the pocket dose is sufficient to ensure $V_{ov,1} < V_{ov,2}$ source to pocket tunneling will overshadow any source edge tunneling effects. Intuitively, both these trends agree well with the offset simulations in the previous section. The significant piece of information gained from this exercise is that the lateral abruptness of the pocket is another critical design parameter for the gTFET.

3.5.3 Dependence on Pocket Length

One of the significant advantages of the gTFET is uniform generation rate occurring across the length of the pocket. This suggests that increasing the pocket length will increase the drive current. In the tunneling limited regime the relationship would be exactly one to one. Doubling the pocket length would double the drive current of the gTFET. In actuality, transport effects complicate the situation. Figure 3.24(a) shows a p-channel gTFET simulation with varying pocket length for fixed overdrive voltage. Above approximately 100 nm pocket length the gTFET drive current saturates. This places an upper limit on the maximum useful pocket length. When compared to the general TFET of Figure 3.1, nearly 40X increase in current is possible with 100 nm gTFET.

However, a 100 nm pocket length gTFET poses a problem for integration density and planar footprint. Since state of the art CMOS is approaching the 22 nm node, a transistor with physical gate length is nearly 5X larger is troubling. One solution is to take advantage of non-planarity in such a manner to allow for large pocket length with small planar gate footprint. One such design is shown in Figure 3.24(b), where the gate is buried within a deep trench. The source and pocket is formed along the trench walls. A large pocket length with smaller planar dimension is possible with this particular design.

While increasing the pocket length increases the gTFET drive current, it is not entirely obvious if this is negated by increased gate capacitance. For fixed gate length of 60 nm various pocket length p-channel gTFETs are simulated. A useful figure of merit which takes into account gate capacities is the CV/I_{on} self delay. However, it is not entirely obvious how to calculate the capacitance C since it will be non-linear with gate voltage. A more useful metric is $\Delta Q_{gate}/I_{on}$, where $\Delta Q_{gate} = Q_{gate}(V_{gs} = 0) - Q_{gate}(V_{gs} = V_{dd})$. This is the amount of charge that must be deposited or removed from the gate to switch the transistor on or off. The I_{on} is the drive current of the gTFET in the on state. From Figure 3.25, it is seen that increasing pocket length still improves the self delay figure of merit but begins to saturate above a certain length. For the 10 nm to 20 nm pocket length case the delay is reduced approximately by 1/2 as expected.

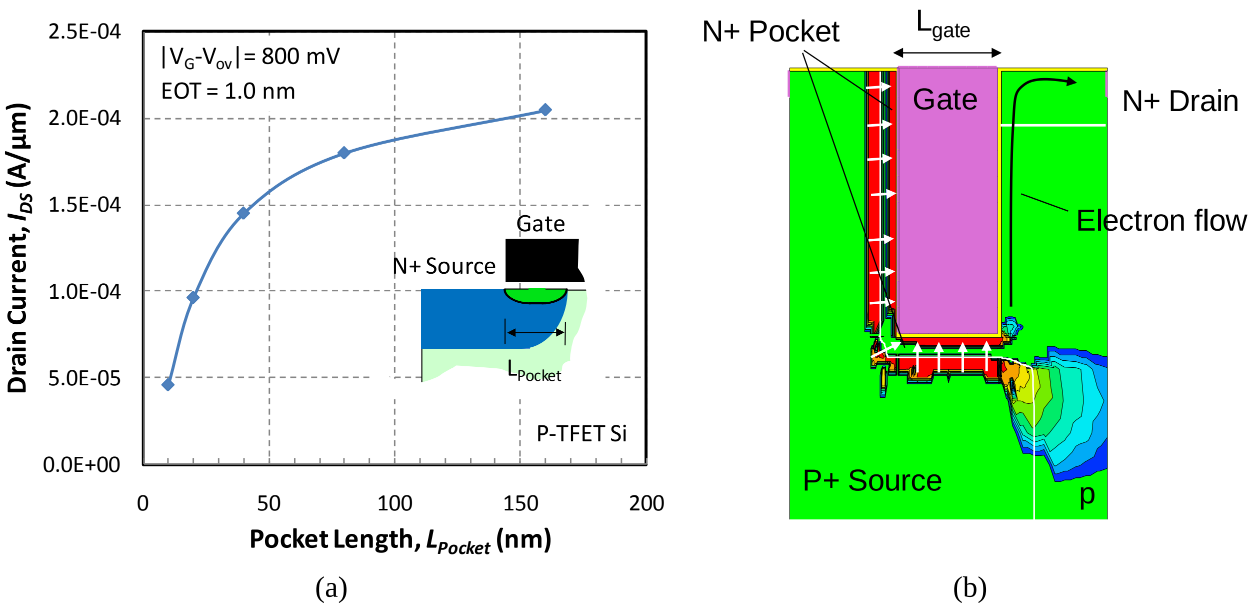


Figure 3.24: (a) p-channel gTFET simulation of drive current vs. pocket length for fixed overdrive. Above 100 nm current begins to saturate resulting from transport effects. (b) Proposed gTFET design that takes advantage of non-planarity to have minimal *L_{gate}* footprint with maximum *L_{pocket}*.

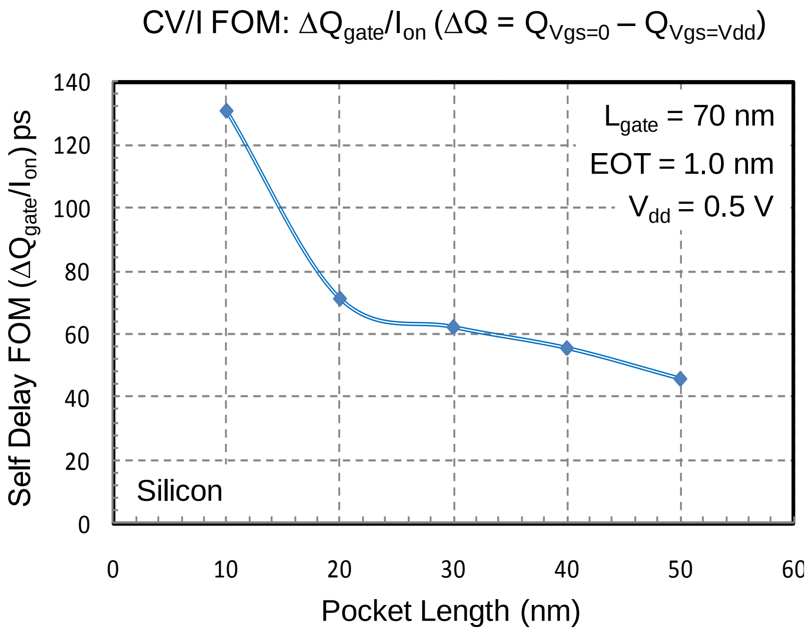


Figure 3.25: Calculation of gTFET self delay metric vs. pocket length for fixed gate length. Large pocket length does improve delay up to certain length.

3.5.4 Comparison of gTFET Input and Output Capacitance with MOSFET

In the previous section it is shown that despite the likely increase in capacitance with increased pocket length, the overall *CV/I* self delay metric still improves up to a certain pocket length. In this section, it is useful to compare the input and output capacitances of a gTFET with that of an identical geometry MOSFET. TFETs in general and especially gTFET (with large source overlap) are expected to have larger capacitance values than the MOSFET. An approximate examination is made into the relative capacitance values of gTFET and MOSFET via simulation.

The input capacitance is the capacitance presented at the gate of the transistor to the output of the previous stage. Since for an MOS system capacitance is non-linear, charge on the transistor gate electrode provides a better measure of "effective" or "average" capacitance. A transistor which requires more gate charge to turn "on", which is defined as $V_{gs} = V_{dd}$, inherently has more capacitance. A fixed gate length of 60 nm is used throughout all simulations. The length of the pocket is varied for a p-channel gTFET from 40 nm to 10 nm for this fixed gate length value. A 60 nm pMOSFET is also simulated as a comparison with threshold voltage adjusted to match that of the gTFET. From Figure 3.26 it can be seen that as pocket length is increased (amount of source overlap) with fixed gate length of 60 nm, the ΔQ_{gate} or effective input capacitance is increased. Compared to the pMOSET reference, the 40 nm pocket length gTFET has approximately 2.3X larger input capacitance for same physical gate length.

The output capacitance is the capacitance presented at the drain of the transistor. This is sometimes referred to as the self capacitance since it "self loads" the transistor. Output capacitance is important when calculating unloaded delay of logic gates, i.e. a transistor whose own capacitance is presented at the load. For this case, a transient simulation is performed on both the identical gate length pMOSFET and gTFET. A simple expression for the unloaded delay is $t_{delay} = 0.69 R_{switch}\left(I_{on}\right) C_{output}$. The pMOSFET threshold voltage is adjusted to equalize the on current with gTFET such that R_{switch} is approximately the same. The output or drain of both transistors is pre-charged initially to $-V_{dd}$ before switching the gate to $-V_{dd}$ to discharge the output. Any difference in the delay time is to first order related to differences in output capacitance between the two devices. As shown in Figure 3.27, the output capacitance of the 40 nm pocket length gTFET is 1.6X larger than pMOSFET.

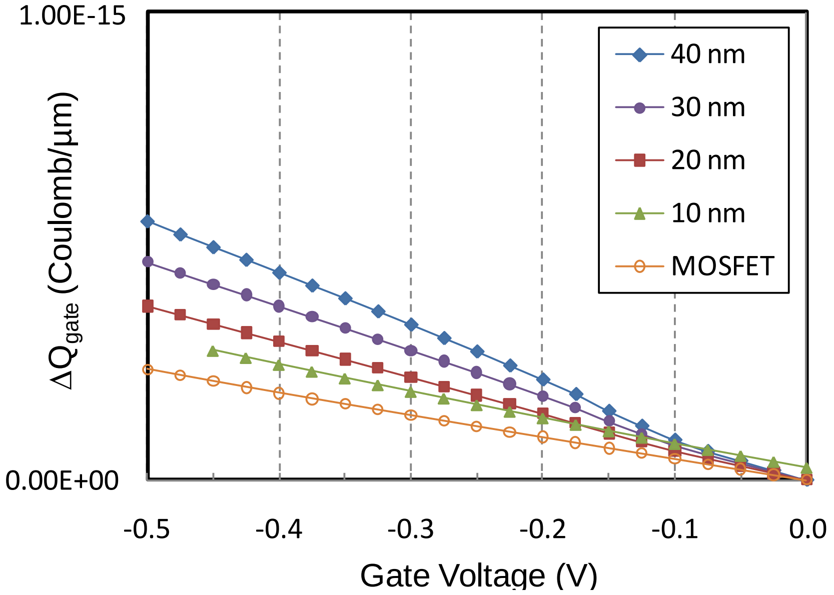


Figure 3.26: Examination of input capacitance of gFET compared to same L_{gate} MOSFET. Charge on gate electrode is plotted vs. gate voltage. On average a 40 nm pocket length gTFET requires 2.3X more gate charge to switch "on". On average, its input capacitance is 2.3X larger.

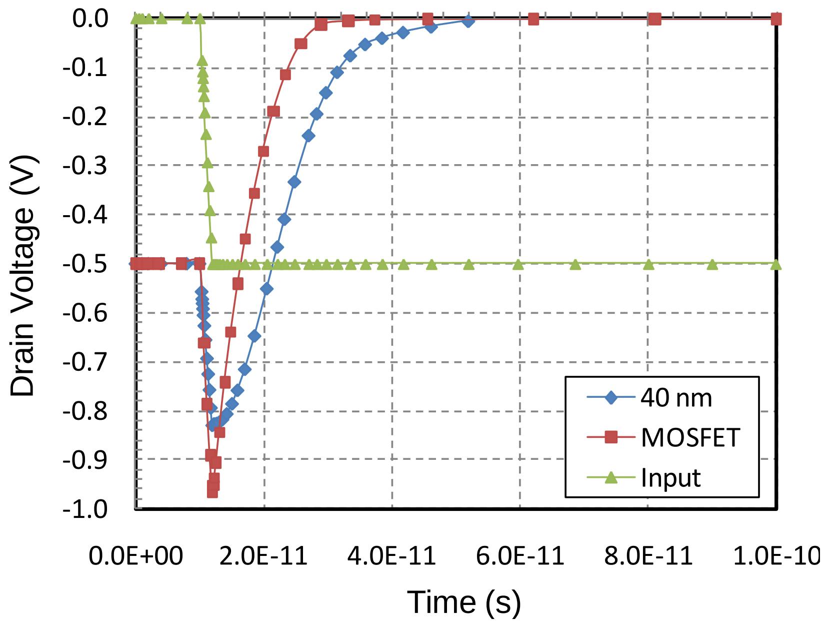


Figure 3.27: Transient simulation of unloaded delay of single p-channel gTFET and MOSFET with drain pre-charged to –V_{dd}. The I_{on} is adjusted to be identical for both devices. The difference in delay is accounted for by output capacitance, which is 1.6X larger for 40 nm gTFET.

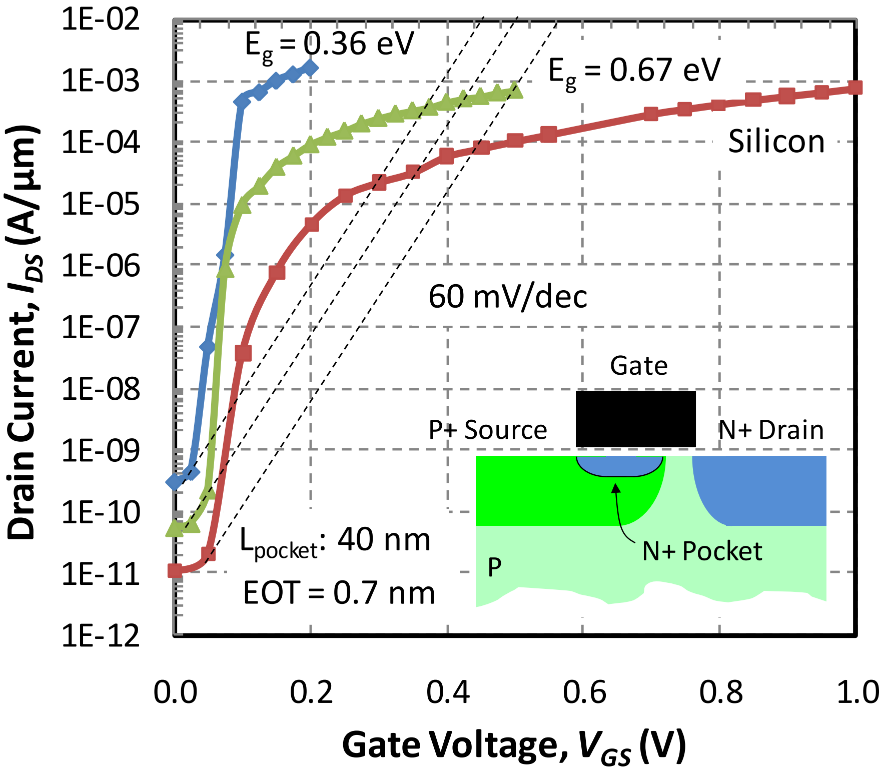


Figure 3.28: Optimized gTFET design for different band gap values. E_g of 0.67 eV corresponds to germanium and 0.36 eV to that of InAs. The tunneling coefficients are scaled by band gap dependence only in this simulation. Tuning the energy gap provides a means of supply voltage reduction.

3.5.5 Optimized gTFETs

Figure 3.28 is the end result of taking into account the importance of all relevant design parameters discussed in the previous section, i.e. large pocket dose, shallow pocket junction, perfect pocket and source lateral alignment, abrupt pocket lateral profile, and large pocket length. The design is very aggressive with pocket junction depth of 2 nm and effective oxide thickness of 0.7 nm. A pocket length of 40 nm is assumed. For the silicon gTFET a supply voltage of 1 V is required to achieve a drive current of approximately 600 µA/µm. By switching the band gap to 0.67 eV, which corresponds to that of germanium, this same level of current is reachable in 0.5 V. In the last curve the band gap is specified as 0.36 eV corresponding to that of InAs. Supply voltage of 200 mV is possible in this design. However, some word of caution must be mentioned. The tunneling coefficients *A* and *B* are scaled only by their band gap dependence from the silicon calibrated values. The change of effective mass is not taken into account. The 0.36 eV which is supposed to be representative of InAs may be largely over predicting the current because of the well known small electron effective mass and DOS issues of InAs. The 0.67 eV and 0.36 eV simulations were also calculated from post process but corrected with transport effects to first order by multiplication of a transport degradation factor as discussed in section 3.3.6 because of non convergence issues. The oxide thickness of 0.7 nm is also considered very aggressive. Therefore, although the quantitative accuracy of the drive current values is in question, the optimized gTFETs of Figure 3.28 are very exciting. Steep “sudden overlap” swing is seen over many decades. When compared to the 60 mV/decade line, the optimized gTFET outperforms. The challenges in fabrication of a well design gTFET are numerous and a potentially limiting showstopper. However, the simulation results are so impressive that some attempt or attempts must be made to realize this device.

3.6 References

[3.1] W. M. Reddick, G. A. Amaratunga, "Silicon surface tunnel transistor," Applied Physics Letters, vol.67, no.4, pp.494-496, July 1995.

[3.2] C. Aydin, A. Zaslavsky, S. Luryi, S. Cristoloveanu, D. Mariolle, D. Fraboulet, S. Deleonibus, “Lateral interband tunneling transistor in silicon-on-insulator," Applied Physics Letters , vol.84, no.10, pp.1780-1782, March 2004.

[3.3] K. K. Bhuwalka, M. Born, M. Schindler, M. Schmidt, T. Sulima and I. Eisele, “P-channel tunnel field-effect transistors down to Sub-50 nm channel lengths,” Japanese Journal of Applied Physics,vol.45, pp.3106-3109, 2006.

[3.4] V. Nagavarapu, R. Jhaveri, J.C.S. Woo, "The tunnel source (PNPN) n-MOSFET: A novel high performance transistor," Electron Devices, IEEE Transactions on, vol.55, no.4, pp.1013-1019, April 2008.

[3.5] C. L. Royer and F. Mayer, “Exhaustive Experimental Study of Tunnel Field Effect Transistors from Materials to Architecture”, International Conference on Ultimate Integration of Silicon, pp.53-56, March 2009.

[3.6] O.M. Nayfeh, C.N. Chleirigh, J. Hennessy, L. Gomez, J. L. Hoyt, D.A. Antoniadis, “Design of Tunneling Field-Effect Transistors Using Strained-Silicon/Strained-Germanium Type II Staggered Heterojunctions,” Electron Device Letters, vol. 29, no.9, Sept 2008.

[3.7] C. Hu, "Green transistor as a solution to the IC power crisis," Solid-State and Integrated-Circuit Technology ICSICT 9th International Conference on, pp.16-20, 2008.

[3.8] A. Bowonder, P. Patel, K. Jeon, J. Oh, P. Majhi, H. H. Tseng, C. Hu, "Low-voltage green transistor using ultra shallow junction and hetero-tunneling," International Workshop on Junction Technology, pp.93-96, May 2008.

[3.9] P. Patel, K. Jeon, A. Bowonder, C. Hu, “A Low Voltage Steep Turn-Off Tunnel Transistor Design,” International Conference on Simulation of Semiconductor Processes and Devices, pp.1-4, Sept. 2009.

[3.10] T. Y. Chan, J. Chen, P. Ko, C. Hu, “The impact of gate-induced drain leakage current on MOSFET scaling”, International Electron Devices Meeting, Vol.33, pp: 718-721, 1987.

[3.11] Y. Tsividis, Operation and Modeling of The MOS Transistor: 2nd Edition, p.95, Oxford University Press, New York, 1999.

[3.12] F. Pregaldiny, C. Lallement, R. Langevelde, D. Mathiot, “An advanced explicit surface potential model physically accounting for the quantization effects in deep-submicron MOSFETs,” Solid-State Electronics, vol.48, pp.427-435, 2004.

[3.13] H.J. Wann, P.K. Ko, C. Hu, “Gate-induced band-to-band tunneling leakage current in LDD MOSFETs,” International Electron Device Meeting, pp.6.5.1-6.5.4, 1992.

[3.14] TFET measurement data from Anupama Bowonder.

[3.15] F. Mayer, C. Le Royer, J.-F. Damlencourt, K. Romanjek, F. Andrieu, C. Tabone, B. Previtali, S. Deleonibus, "Impact of SOI, $Si_{1-x}Ge_x$OI and GeOI substrates on CMOS compatible Tunnel FET performance," International Electron Devices Meeting, pp.1-5, Dec. 2008.

Chapter 4: Fabrication of Green Tunnel Field Effect Transistors

4.1 Introduction

Simulation results of Chapter 3 demonstrate that the gTFET can achieve very steep swing over many decades of current by careful dopant engineering using ultra shallow "pockets" of charge. Fabricating this device, however, is a challenge. A very heavily doped, ultra shallow, and near perfect lateral profile termination is needed to "see" the "sudden overlap" steep swing in the simulation results. In this chapter the various fabrication attempts of the gTFET are discussed. Initial silicon gTFET measurements are presented and analyzed.

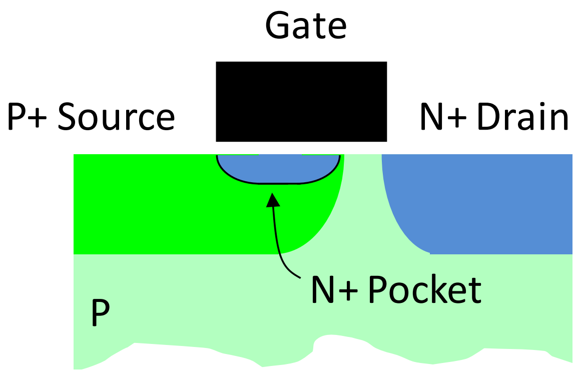


Figure 4.1: The gTFET is a low voltage transistor that poses significant fabrication challenges especially concerning pocket formation.

4.2 Fabrication of gTFET at Sematech

The simulations of the gTFET design space have shown a necessity for excellent pocket doping profile control, i.e., high dose, ultra shallow junction, near perfect lateral profile termination. This makes the fabrication of the gTFET extremely challenging. One of our main collaboration partners for this project under the DARPA STEEP research grant was Sematech. Sematech houses advanced process modules and capability that is not available in the UC Berkeley Microlab. Ultra low energy implantation, millisecond flash annealing, and advanced and mature high-k dielectric gate stack are a few of the many process modules that are critical for fabrication of a well designed gTFET. I spent one year on site at Sematech working with integration engineers to develop a gTFET process flow, running various gTFET experimental lots in their advanced 200 mm clean room, and fully characterizing the fabricated devices. Many of the exact process details, i.e. recipe conditions, are not described in this chapter for the Sematech experiments. The important split conditions, however, are noted. The Berkeley Microlab fabrication runs to be discussed in the following sections, however, are described in full detail in the Appendix at the end of this chapter.

The "most ideal" way to fabricate the gTFET would be to use selective doped epitaxial growth of silicon to form both source and pocket. This would result in the most abrupt pocket and source lateral doping profile. However, this capability did not exist at the time the experiments were run and is still currently under development. The next most obvious choice is to use low energy ion implantation for formation of source and pocket. Once choosing the implant route the reminder of the process steps are self determinant as seen as follows. From the gTFET structure with large source overlap, it becomes obvious that these implantations must be performed prior to gate stack formation. This is known as a "pocket first" process. Also since half of the channel region is heavily doped, growing the gate dielectric by oxidation is questionable. The rate of oxidation has strong dependence on doping concentration. The end result would be a gate oxide that is thicker under the pocket and thinner in the lighter doped channel region. This requires that the gate dielectric be deposited rather than grown. A high-k metal gate stack is the ideal choice. Dopant activation should also minimize any dopant diffusion of the pocket and source profile. A millisecond flash anneal, where wafers can be heated to temperature of approximately 1200° C for millisecond duration, is required. In this process, wafers are heated uniformly to an intermediate temperature then subject to a high intensity short duration flash lamp, which heats the surface to very high temperature for very short time. This process has minimal diffusion. [4.1-4.3]

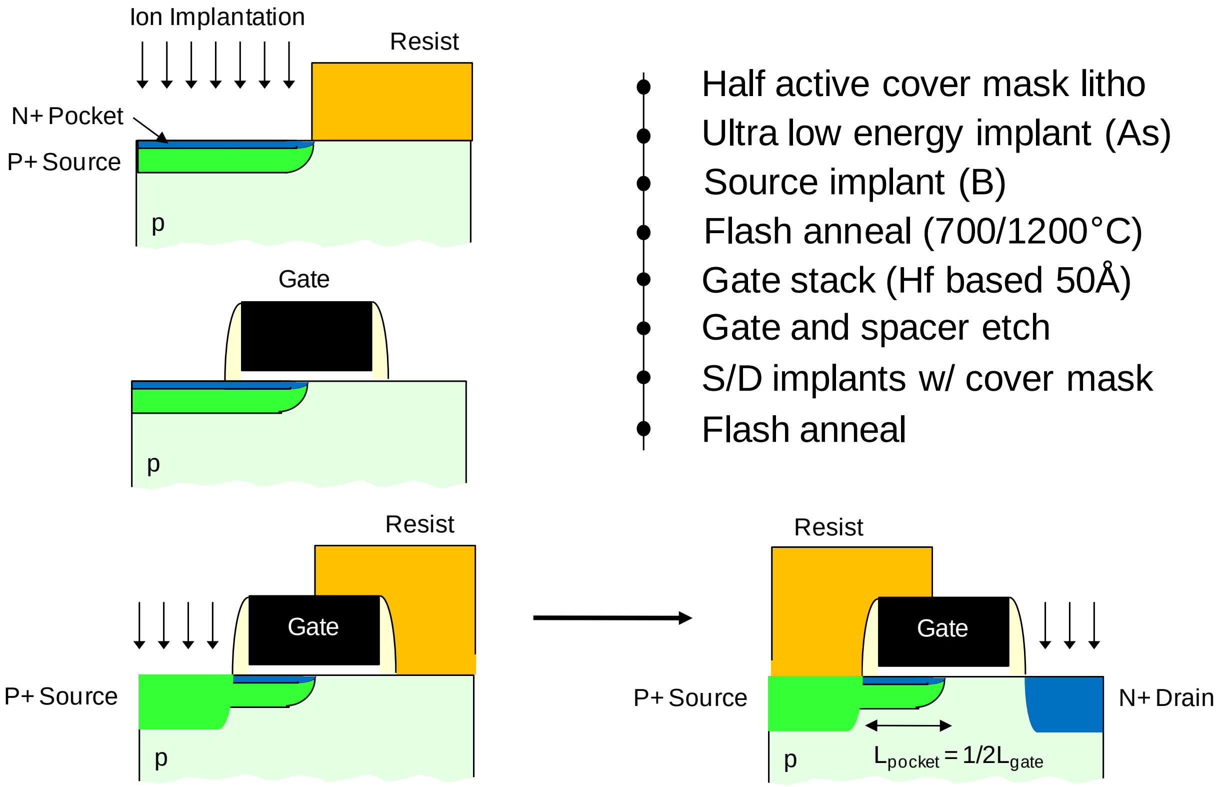

Ion Implantation
Resist
N+ Pocket
P+ Source
p
Gate
p
Half active cover mask litho
Ultra low energy implant (As)
Source implant (B)
Flash anneal (700/1200°C)
Gate stack (Hf based 50Å)
Gate and spacer etch
S/D implants w/ cover mask
Flash anneal
Resist
Gate
P+ Source
p
Resist
Gate
P+ Source
N+ Drain
p
$L_{pocket} = 1/2L_{gate}$

Figure 4.2: Simplified process flow used for all gTFET experiments at Sematech. Ultra low energy implantation is used to define the pocket and the source. Flash annealing is used to activate the dopants with minimal diffusion. A deposited high-k stack is used as the gate dielectric.

The general gTFET process flow used in all Sematech fabricated wafer runs is shown in Figure 4.2, which is a modification of the standard baseline silicon MOSFET flow. The baseline is used for the non-essential steps: cleans, active definition/isolation, gate etch, spacer formation, contact and backend metallization. The starting silicon wafers have active layers that are shallow trench isolated. A half active implant mask is first used to block half the active region. Both As pocket and B/BF$_2$ source are implanted aligned to the mask edge. The wafers are then flash annealed to remove implant damage and to activate the dopant atoms. ALD high-k which is hafnium based is then deposited and subject to some post deposition anneal with nitridation. [4.4-4.8] ALD TaN is deposited as the gate electrode. After standard gate etch and spacer formation, the half active mask is used to block the drain side while implanting the deep P+ source region. The opposite of this mask is used again to cover source while N+ deep drain is implanted. Another flash anneal step is performed to activate the deep source and drain while minimizing diffusion of the pocket and source profiles. The remaining steps are the standard baseline backend contact and metallization.

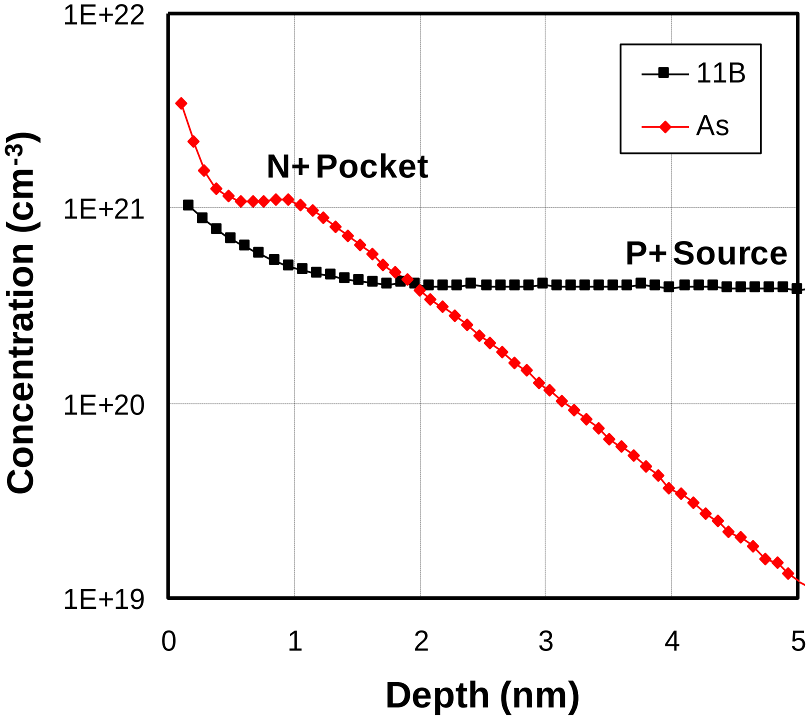


Figure 4.3: SIMS profile of low energy arsenic implantation subject to flash anneal at 1250° C peak temperature. Some artifacts are noticed near the surface. Pocket junction depth of 3 nm is possible if source doping concentration is 1E20 cm^{-3}.

4.2.1 Arsenic Implanted Pocket Module Development

The pocket is the single most critical element of the gTFET. An improperly designed pocket results in significant degradation of the gTFET I_d-V_g curve as shown in Chapter 3. Initial research focus is devoted to developing an implanted arsenic pocket module. Test wafers are blanket implanted with arsenic and BF_2 at ultra low energies. SIMS is used to probe the doping profile along the depth direction within the first few nanometers of the surface. There are known limitations on the resolution of SIMS near the surface of the sample. Some artifacting can be seen within the first nanometer in Figure 4.3. In this sample, arsenic was implanted with 500 eV and 1E14 cm^{-2} dose followed by BF_2 of 5 keV and 2E14 cm^{-2} dose. The sample was then flash annealed at 750° C intermediate temperature with peak of 1250° C. This experiment shows that it is possible to form approximately 3 nm implanted N+ pocket junction if the P+ source doping concentration is at the 1E20 cm^{-3} level.

In order to test the electrical activation and confirm that the pocket is shallow enough to be fully depleted simple CV structures are fabricated. A low energy arsenic implantation is performed through a capacitor dot isolated p-type wafer and flash annealed at the same condition as above. A standard high-k gate stack is deposited on top. Figure 4.4 shows the measured CV characteristic for various dose arsenic implants compared to a control sample. If the N+ pocket is fully depleted the CV curve will shift to the left by an amount proportional to the activated dose.

$$\Delta V_{fb} = \frac{-Q_{dose}}{C_{ox}} \qquad (4.1)$$

Figure 4.4 confirms that the pocket is fully depleted from the 7E13 cm^{-2} case. From knowledge of the C_{ox} and amount of flat band shift an electrically activated dose of approximately 30% is extracted.

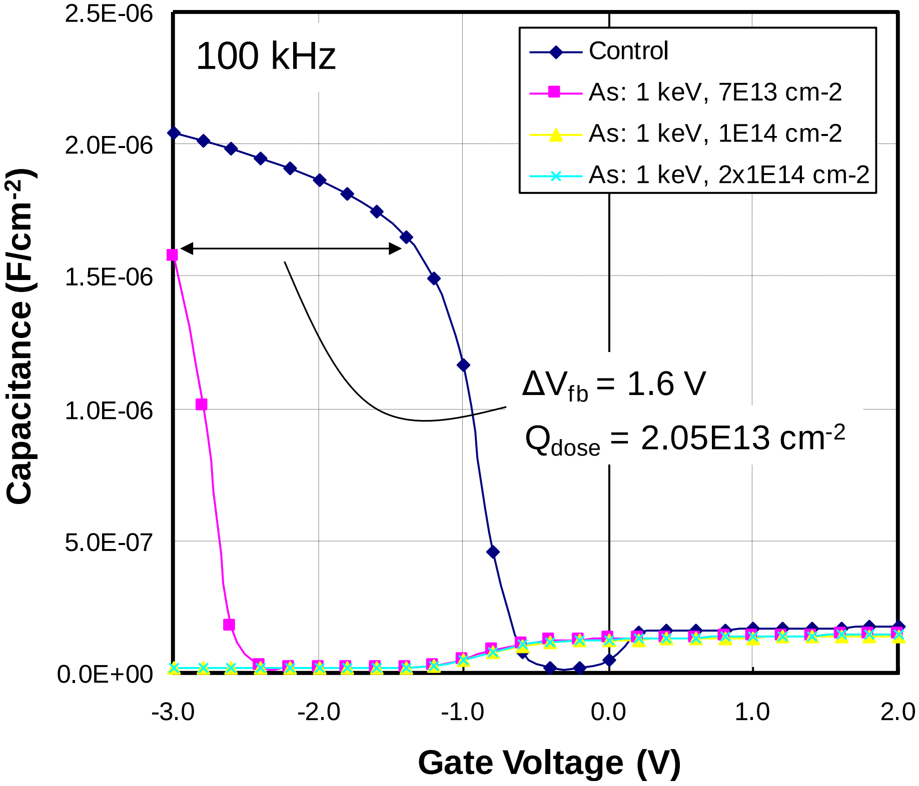


Figure 4.4: CV measurement of capacitors with varying amount of low energy arsenic implantation. An electrically active dose can be extracted from the flat band voltage shift of the curves.

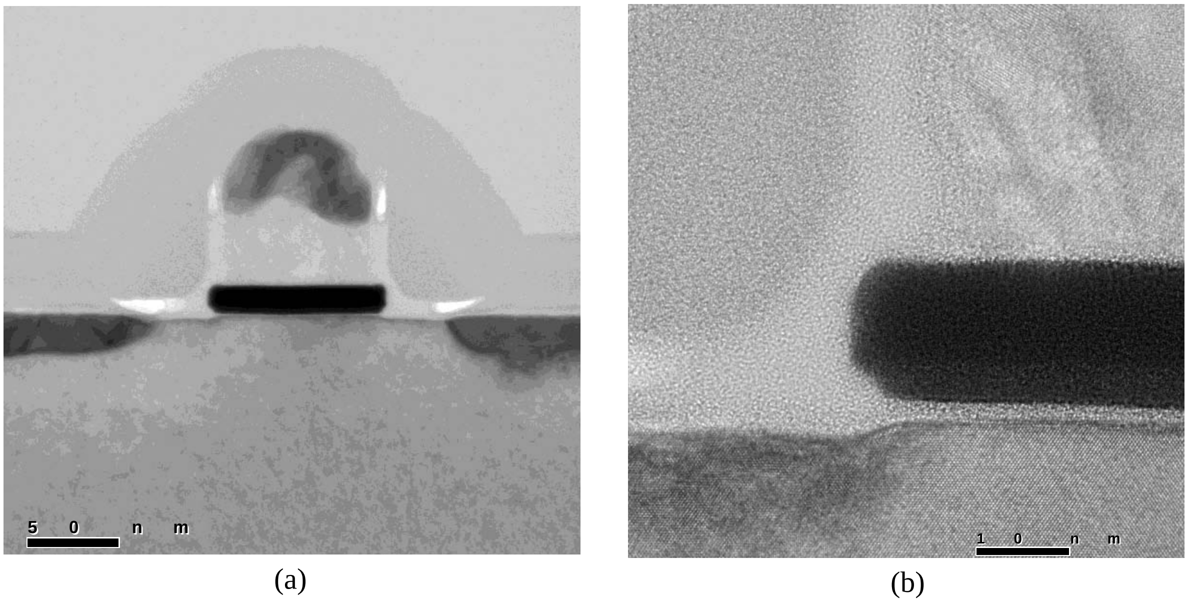


Figure 4.5: (a) X-TEM of the fabricated gTFET. (b) Good crystallinity is observed in the implanted pocket and source region under the gate.

4.2.2 Initial gTFET Fabrication Attempt

An initial few wafers are run through the process flow of Figure 4.2 to debug the process and to discover any possible yield showstoppers. Of particular concern is possible wafer warping and breakage from two flash anneals. For MOSFETs following the baseline flow only a single flash has been attempted. Figure 4.5 shows a cross section TEM of the gTFET after fabrication. Good crystallinity can be observed in the channel region suggesting most of the pocket and source implant damage has been annealed away. 50 Å of ALD HfSiON is used as the gate dielectric for these wafers. The flash anneal conditions are 750° C intermediate and 1250° C peak. A single arsenic pocket and boron source condition is used. (As: 1 keV, 7E13 cm^{-2} and BF$_2$: 3 keV, 1E14 cm^{-2}) The measured gTFET I$_d$-V$_g$ is shown in Figure 4.6. No “sudden overlap” steep swing is seen on these initial gTFETs. Figure 4.7 shows the I$_d$-V$_d$ characteristics, where typical TFET non-linear behavior is seen at the low V$_{ds}$ regime.

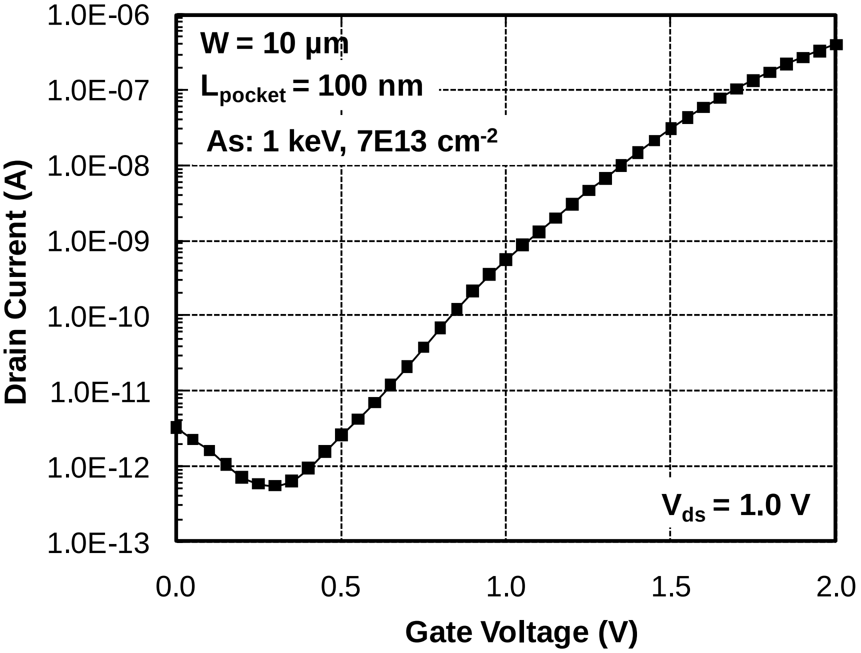


Figure 4.6: Measured gTFET I_d-V_g showing gradual turn on characteristics. No "sudden overlap" swing is observed across any devices in this initial experiment. The source implant condition is BF_2: 3 keV, 1E14 cm^{-2}.

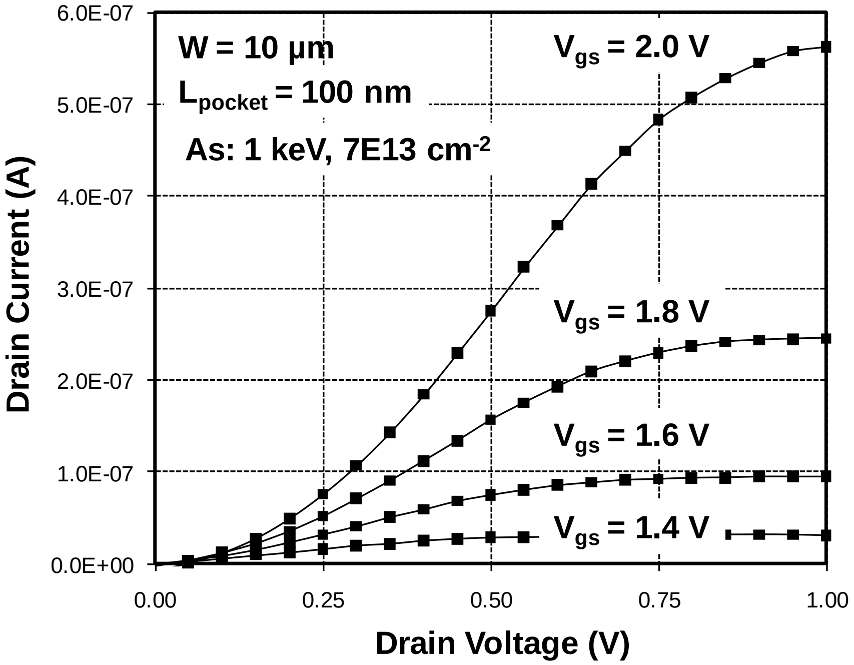


Figure 4.7: Measured I_d-V_d characteristics for fabricated gTFET showing non-linear behavior at low V_{ds} and good output conductance.

4.2.3 gTFET Comparison to Control Wafer

In this set of experiments, a control wafer is included that does not receive the pocket implant. 30 Å of HfO_2 is used as the gate dielectric. The flash anneal condition is 750° C / 1250° C peak. The same flow of Figure 4.2 is used. Two different arsenic implant conditions are compared with the control wafer in Figure 4.8. The I_d-Vg curves are identical to the control but horizontally shifted. This means that for these implant conditions, source edge tunneling and not source to pocket tunneling has determined the turn on characteristics. Since implantation is an imprecise method of forming the N+ pocket, a larger split on both arsenic and boron energy and dose needs to be run.

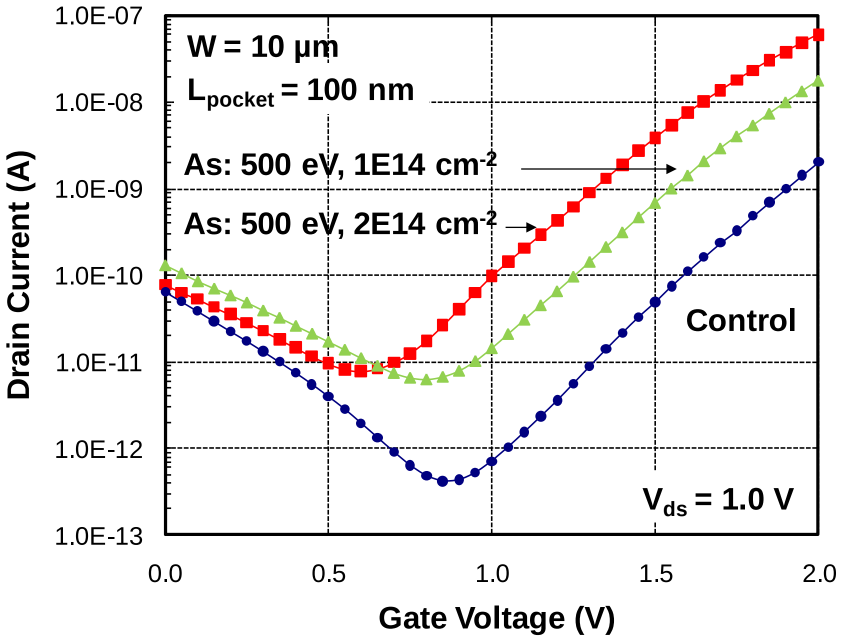


Figure 4.8: Measured gTFET I_d-V_g compared to control device which did not see any arsenic pocket implantation. The I_d-V_g curves are identical with exception of voltage shifts. This suggests source to pocket tunneling is not occurring in these gTFETs. The source implant condition is: BF_2 3 keV, 1.5E14 cm^{-2}

4.2.4 Summary of Expanded Pocket Energy and Dose Variation

For this set of experiments expanded splits on gTFET pocket implant energy and dose as well as source dose are fabricated. The same 50 Å of HfSiON gate stack is used with flash anneal condition of 750° C / 1250° C peak. The process flow remains the same as shown in Figure 4.2. Figure 4.9 shows the summary of various As pocket implant splits with source implant condition held constant. Unlike, the previously shown measurement, some modulation of the swing and overall drive current is seen. The best I_d-V_g characteristic has swing of approximately 100 mV/dec and corresponds to arsenic implant condition of 1 keV and 1E15 cm^{-2} dose. This large amount of implant dose suggests that only a small fraction of the dopants are activated. An interesting split to include in future wafer runs would be an even higher arsenic dose of greater than 1E15 cm^{-2}.

In Figure 4.10, the arsenic implant condition is held constant while boron dose is changed. An optimum boron dose of 1E15 cm^{-2} is observed to produce a swing of approximately 100 mV/dec. When the dose is too low, the source doping concentration is low resulting in poor turn on characteristics. When too high, the pocket doping concentration is compensated also resulting in poor swing. The temperature dependence of the I_d-V_g is shown in Figure 4.11 from 300 K down to 100 K. The swing is plotted vs. drain current and is shown to be independent of temperature. This provides confirmation that these fabricated gTFETs are not accidental MOSFETs, where swing would be proportional to kT. Figure 4.12 provides some confirmation that these gTFETs are dominated entirely by “source edge tunneling”. The drain current is shown to be completely independent of pocket length, indicating that source to pocket tunneling is not occurring in these devices.

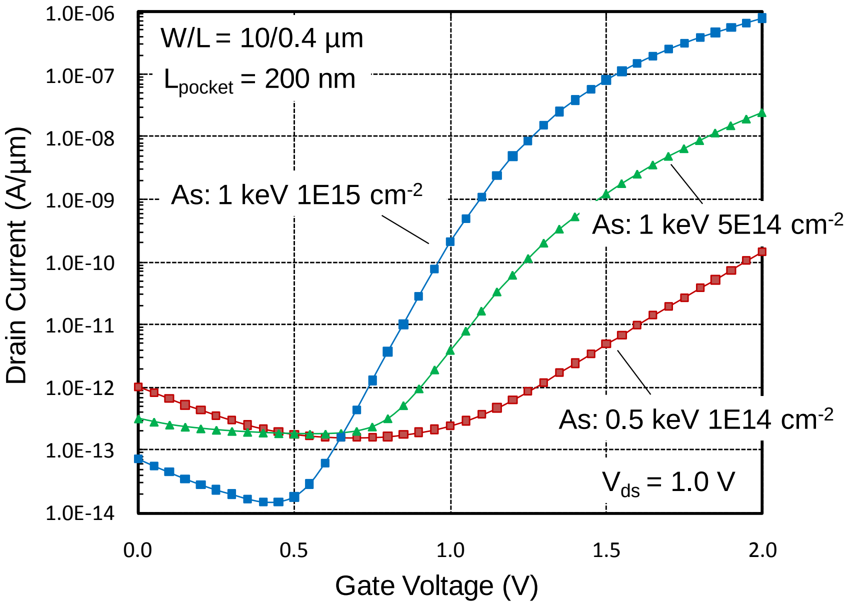


Figure 4.9: Measured gTFET I_d-V_g over wide range of arsenic pocket implant conditions. The source implant is held constant at: 4 keV, 3E15 cm^{-3}. Modulation of swing is seen with increasing pocket implant dose. Best swing is approximately 100 mV/dec.

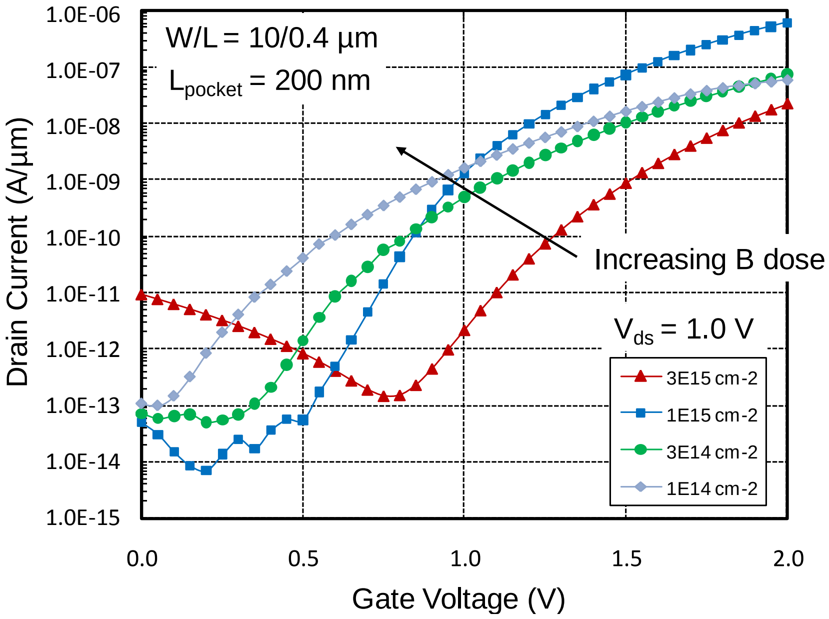


Figure 4.10: Measured I_d-V_g for gTFET with fixed arsenic implant of 1 keV and 5E14 cm^{-2} and varying B source dose. An optimal dose of 1E15 cm^{-2} is seen to produce approximately 100 mV/dec. swing.

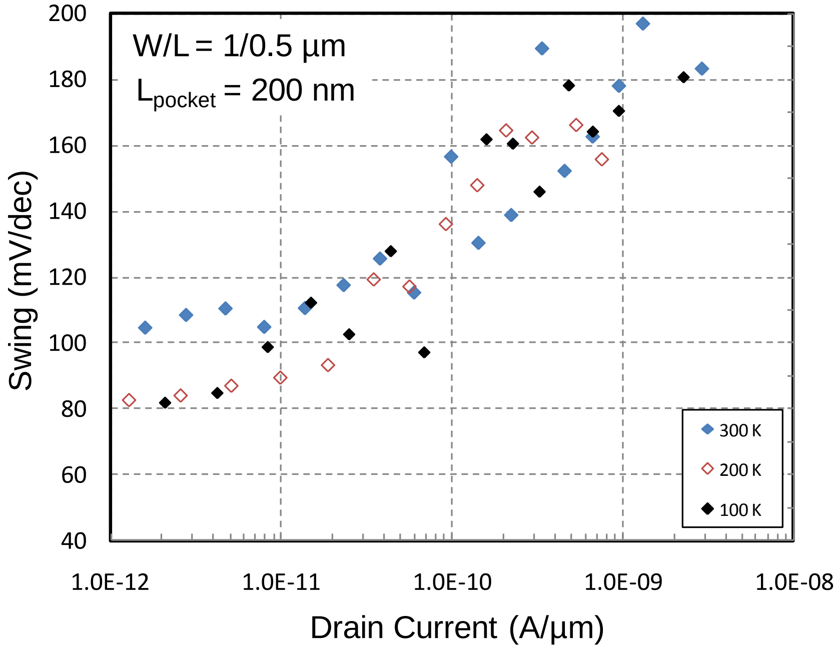


Figure 4.11: Measured swing vs. drain current for fabricated gTFET across various temperature. The swing is not proportional to kT and confirms that the device is not a MOSFET.

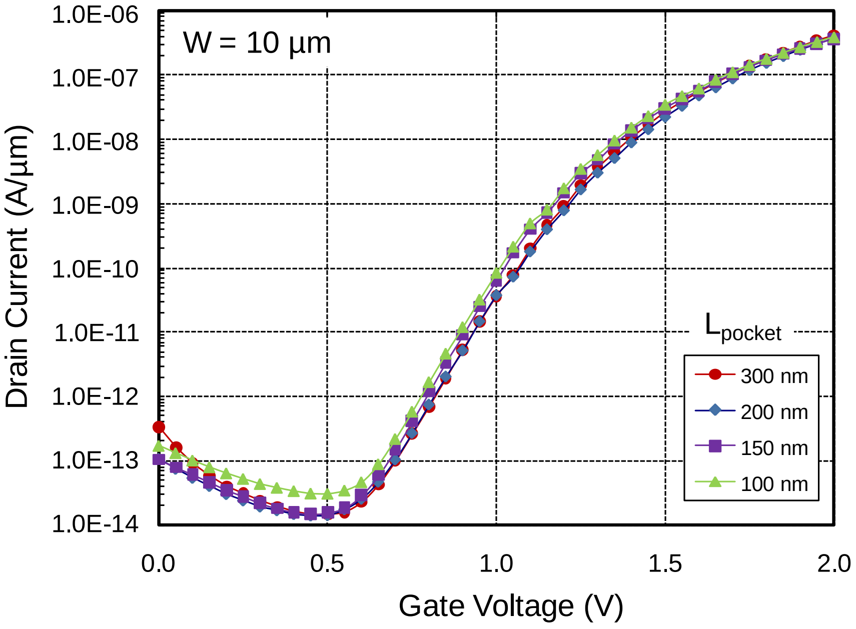


Figure 4.12: Measured gTFET I_d-V_g for various pocket lengths. The insensitivity of the current to pocket length suggests that source to pocket tunneling is not occurring. The tunneling is likely occurring in the source edge region.

4.2.5 Explanations for Lack of "Sudden Overlap"

Thus far none of the gTFET measurements have shown any "sudden overlap effect". The smallest swing that has been measured is 100 mV/dec among all wafers across all energy and dose splits in lots fabricated at Sematech. The most likely explanation is that the gTFET pocket was not properly designed. In Chapter 3 simulations have shown that a poorly designed pocket results in very poor swing or "turn on" characteristics in the gTFET. In particular, the lateral pocket doping profile abruptness is most critical. While low energy implants can permit ultra shallow pocket junction, the lateral straggle of the implant is still non-abrupt. Fundamentally, the ion implantation process is simply not precise enough to form a well designed pocket resulting in poor gTFET demonstration. Alternative means of pocket formation using epitaxial growth need to be explored as will be discussed.

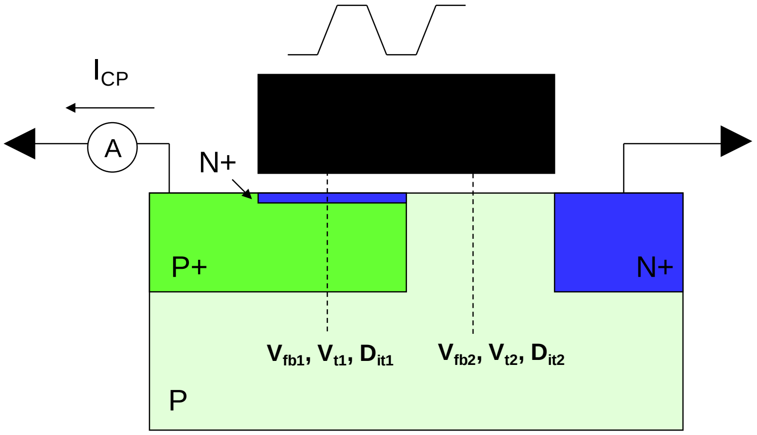

I_{CP}
A
N+
P+
N+
V_{fb1}, V_{t1}, D_{it1}
V_{fb2}, V_{t2}, D_{it2}
P

Figure 4.13: In charge pumping the gate voltage pulse must be large enough to switch between flat band and inversion to sweep the entire range of interface traps within the band gap. Assuming D_{it} from the pocket region is dominating the pulse must be between V_{fb1} and V_{t1}. The theory of this technique is described in detail in [4.9-4.10].

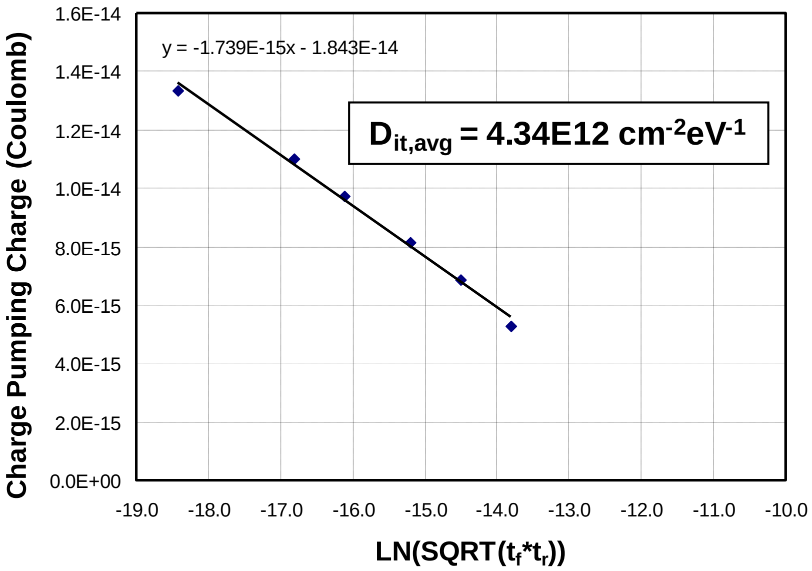


Figure 4.14: The average D_{it} can be extracted by varying the rise and fall times of the pulse as described in detail in [4.9]. Large average D_{it} of 4E12 $cm^{-2}eV^{-1}$ is extracted in the pocket region suggesting poor gate dielectric interface quality as suspected.

In addition, it is suspected that the interface quality (very high trap density D_{it}) between a very heavily doped semiconductor and gate dielectric is very poor. There is no prior research on this subject to confirm this suspicion, since MOSFETs typically involve channel doping in the 1E18 cm^{-3} range at most. If the D_{it} is very large over the pocket region, the potential will be essentially "pinned" in this region. This means that the overlap condition for source to pocket tunneling will never occur. In other words, the turn on characteristic will be dominated by source edge tunneling.

Performing charge pumping on the fabricated gTFETs is one way to extract D_{it} information. Figure 4.13 describes the measurement test setup. The gTFET is a non-standard structure for charge pumping analysis because half the channel is heavily doped while the other half is lightly doped. It is assumed that the D_{it} over the heavily doped pocket region contributes to most of the measured charge pumping current. Average D_{it} values of 4E12 $cm^{-2}eV^{-1}$ are obtained for the interface between gate dielectric and pocket. This result provides some confirmation that fabricated gTFETs suffer from poor interface properties in the pocket regions.

4.3 Pocket Last gTFET Fabrication

The previous section has highlighted the challenges of good dielectric interface formation for gTFET. Depositing dielectric on heavily doped surface results in high D_{it}, causing tunneling to occur only in the undesired source edge region. The potential in the pocket region is effectively "pinned". One possible solution is to form a good quality gate dielectric stack first then implant to form the source and pocket region. This "pocket last" process flow might alleviate some of the interface quality concerns. This experiment was fabricated in Berkeley Microlab following a process flow shown in Figure 4.15.

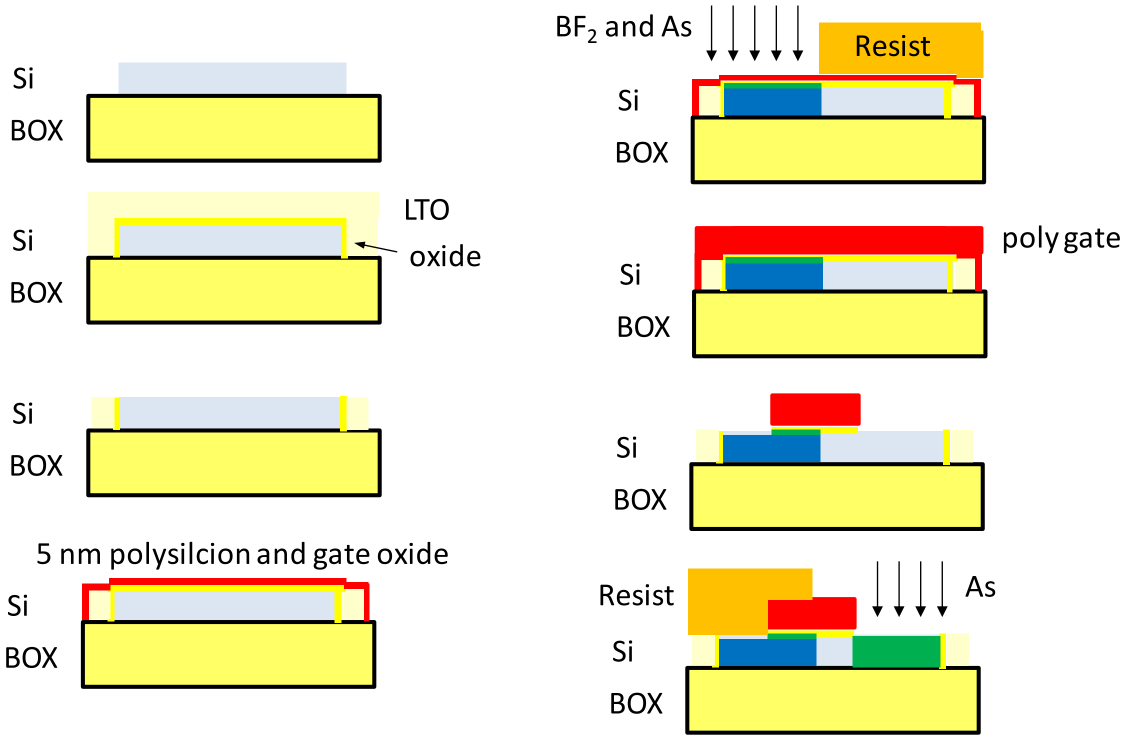


Figure 4.15: Simplified process flow for "pocket last" experiment. Good quality thermal oxide is grown on silicon first. The pocket implant is formed after partial gate dielectric deposition, where arsenic is implanted through the gate oxide.

4.3.1 Process Details

The starting wafers are 6” p-type 100 nm body silicon on insulator. Dry oxidation followed by HF dip is used to thin the silicon body to 40 nm. Active lithography is performed and the silicon is reactive ion etched to the buried oxide to form islands of active (silicon mesas). Gate oxidation is performed at 800° C for 3 minutes to grow approximately 2.8 nm of oxide. The wafers are transferred immediately to a CVD furnace to deposit 5 nm of in-situ phosphorus doped polysilicon as gate electrode. A half active covering mask is used to block the drain while low energy pocket and source are formed by implantation through the initial gate stack. After HF pre-clean more polysilicon is deposited on top of the existent layer to a final thickness of 1000 Å. Gate lithography is performed and reactive ion etching is used to etch the polysilicon. Note that because of the step height between the buried oxide and silicon active mesa, a gate stringer forms during the gate etch. Additional over etch is needed to completely remove the stringers around the mesa. The selectivity to the underlying gate oxide in the source and drain regions is a problem. The control wafer (did not receive arsenic pocket implant) was lost in this process step as a result of completely etching away source and drain. Figure 4.16 shows the top down SEM after successful gate etch. Afterwards, 500 Å of oxide is deposited and reactive ion etched to form a gate aligned spacer. The drain and source blocking mask are then applied for deep source and deep drain implantation respectively. The wafers are capped with 1500 Å of LTO from CVD furnace. The dopants are activated with 1020° C spike anneal for 2 s. Contact lithography is performed and the openings are reactive ion etched. The devices are directly probable and no metallization is performed. Standard forming gas anneal at 400° C for 30 mins in H_2 ambient is the last processing step.

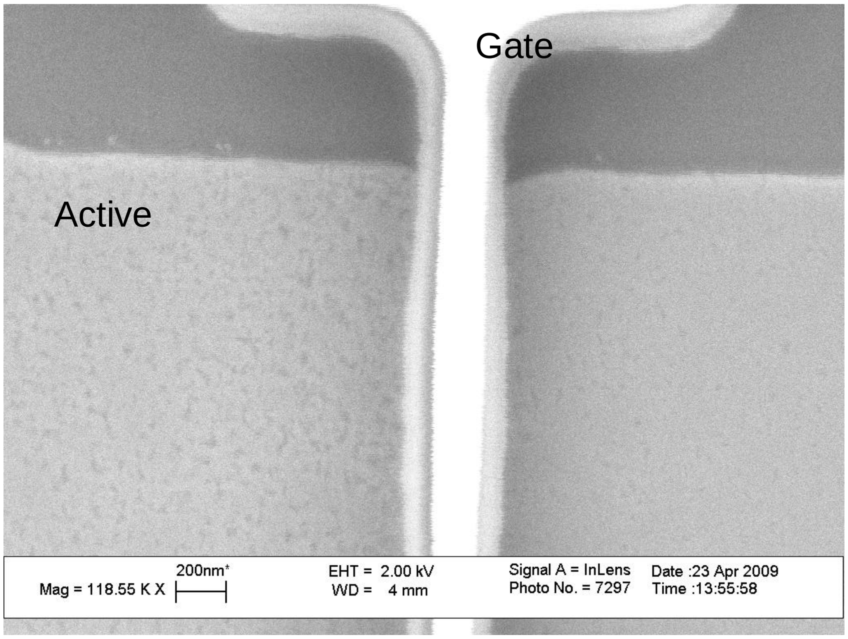


Figure 4.16: SEM image of the active and gate layer after polysilicon gate etch and clean.

4.3.2 Measurement Results

Unfortunately, all fabricated devices suffered from catastrophic gate leakage as shown in Figure 4.17. This I_d-V_g curve is representative of all the wafers across the energy and dose splits. The gate to drain leakage masks any band-to-band tunneling that may be occurring in the semiconductor. The likely cause is damage from the pocket and source implant through the gate oxide. Although implant through gate dielectric has been reported in literature [4.11], the energy of implantation is much lower for these wafers. Although some amount of gate leakage is tolerable, the excessively large amount present in these devices suggests that the "pocket last" process should be avoided in future experiments. Currently there is no obvious way of forming a "pocket" after the gate stack is deposited without damage to the gate dielectric. Future experiments all revert back to the "pocket first" approach used at Sematech.

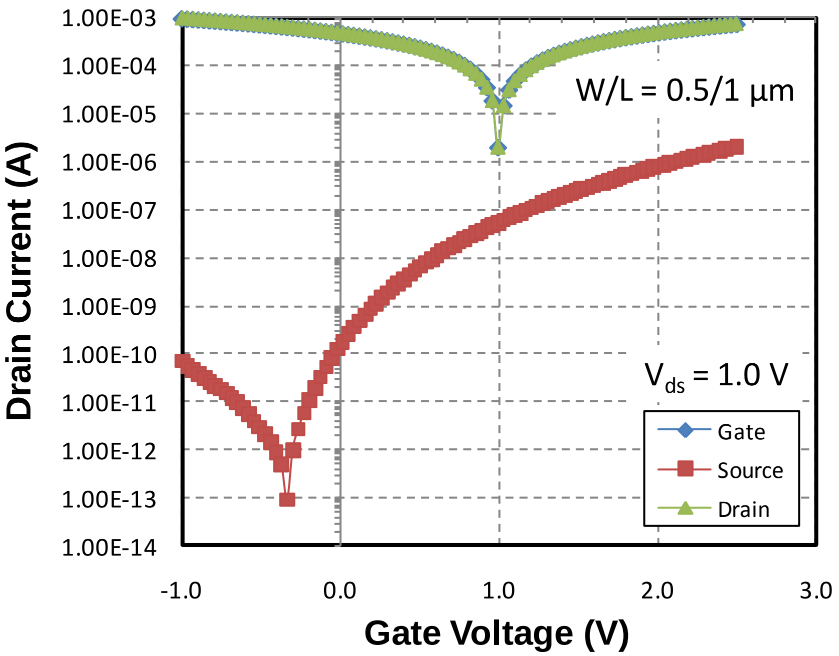


Figure 4.17: Measured "pocket last" gTFET I_d-V_g showing very large gate leakage between gate and drain. This is likely resulting from damage from the pocket implant. No band-to-band tunneling is observed in any of the devices across all wafers. The device shown had As: 5 keV, 2E14 cm^{-2} and BF_2: 5 keV, 3E14 cm^{-2}.

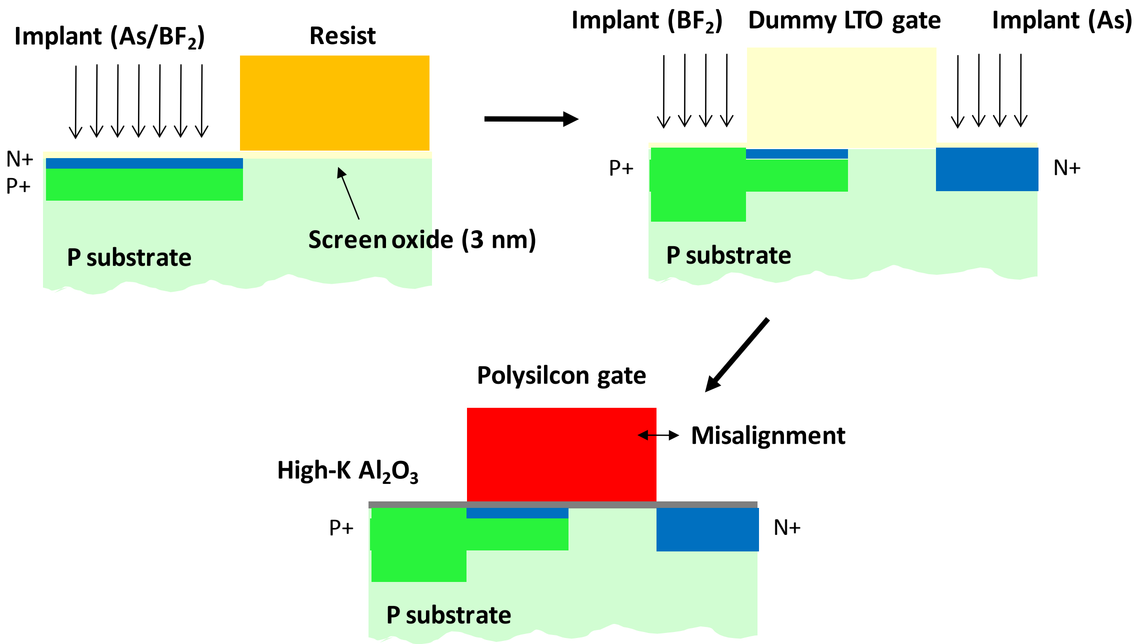


Figure 4.18: Simplified process flow of the modified pocket first experiment. Screen oxide is grown before the pocket and source implant is performed. A dummy oxide gate is used to form deep source and drain. The actual gate is deposited later, which can be potentially misaligned from the implants as shown in the figure above. This process avoids large thermal budget and implant related damage once the high-k dielectric is deposited.

4.4 Modified Pocket First gTFET Fabrication

In this process the concern of sufficient pocket dose near the wafer surface is addressed. It is a known problem that achieving high activated dose for low energy implants is challenging. At ultra low energy, sputtering and removal of the first few nanometers of the silicon surface during the implantation process self limits the incorporated dose. [4.12-4.13] In this experiment, a thin layer of oxide is grown before the pocket and source implant to help incorporate more pocket dose. Also, more of the pocket dose design space is explored in hopes of achieving "sudden overlap" effect. Spike anneal is used for activation, which is non ideal in terms of dopant diffusion, but may result in better surface damage removal and potentially improved interface. Most importantly, this experiment is carried out in the pioneering spirit of fabricating a new exciting device and learning as much as possible from the results to advance knowledge in this field.

The wafers were fabricated in Berkeley Microlab following the process flow shown in Figure 4.18. An isolation free ring FET mask is used for quick turnaround of experiments. The ring FET also avoids problems with gate stringer around mesa isolated SOI that has plagued the previous "pocket last" experiment. All implantation is done prior to gate stack deposition and then spike annealed for activation, i.e. source, pocket, deep source, deep drain. This is to ensure the deposited gate dielectric Al_2O_3 does not crystallize during the thermal budget of the anneal. The Al_2O_3 is also not subject to any edge implant damage as a result of this process.

4.4.1 Process Details

A table containing all process details of this experiment is shown in the appendix of this chapter. Bulk p-type wafers are the starting material in this process. A 3 min 800 °C dry oxidation is performed to grow approximately 3 nm of capping oxide. A half active implant mask is used to cover the drain side while low energy arsenic is implanted to form pocket. BF_2 is implanted at 10 keV 5E14 cm^{-2} for formation of source. Afterwards, 1000 Å of LTO is deposited in an LPCVD furnace. Gate lithography is performed and the oxide is reactive ion etched leaving approximately 100-150 Å remaining in the unprotected regions to form a dummy gate. Source and drain blocking lithography is performed with BF_2 and As implantation respectively for deep source and drain formation. Spike anneal at 1040 °C peak and 1 second duration is performed for activation of dopants. Afterwards, the dummy oxide gate is removed in HF bath in preparation for gate stack formation. ALD Al_2O_3 of 50 Å is deposited (50 cycles) then immediately loaded into CVD furnace for deposition of 1000 Å of in situ phosphorus doped poly silicon as gate material. Gate lithography is performed again and the gate stack is reactive ion etched. Since the wafer is planar, no gate stringer removal is required. There is some misalignment between the actual gate and the dummy oxide gate determined by the alignment tolerance of the lithography tool. This results in either overlap or underlap (no overlap) of the deep drain. 1500 Å of LTO is deposited and contact lithography is performed. Contact openings are reactive ion etched and Al 2% silicon is deposited for metallization. Forming gas anneal is performed at 400 °C for 30 min in H_2 ambient.

4.4.2 Measurement Results

Figure 4.19 shows the p-channel I_d-V_g from the control wafer, which did not receive arsenic pocket implant. The subthreshold swing is not very steep, as can be expected since this is the control sample. The I_d-V_d curve of Figure 4.20 shows strong non-linearity at low drain bias and good saturation characteristics as seen in most TFETs. The impact of FGA on the IV characteristic is shown in Figure 4.21, which is minimal. The n-channel control device, unfortunately, suffers from gate leakage issues that overshadow any potential band-to-band tunneling that may be occurring in the semiconductor as seen in Figure 4.22. The amount of gate leakage is surprising since the Al_2O_3 was not exposed to either high temperature or implant damage. This is a problem because the wafers which have received a pocket implant have no equivalent n-channel control reference.

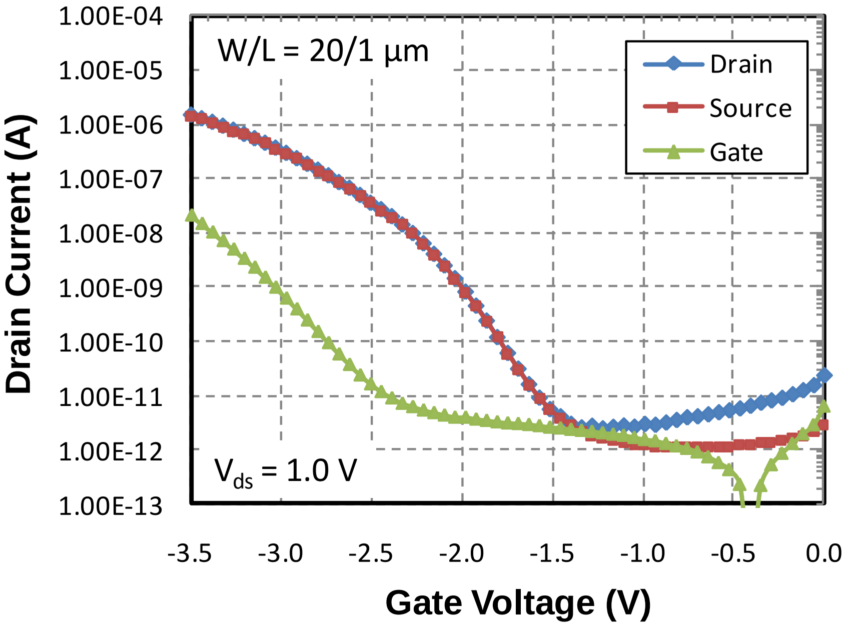


Figure 4.19: Measured I_d-V_g of p-channel control device, which did not receive an arsenic pocket implant. The overlap voltage is shifted by approximately 1 V in the p-channel device because N+ poly is used for the gate.

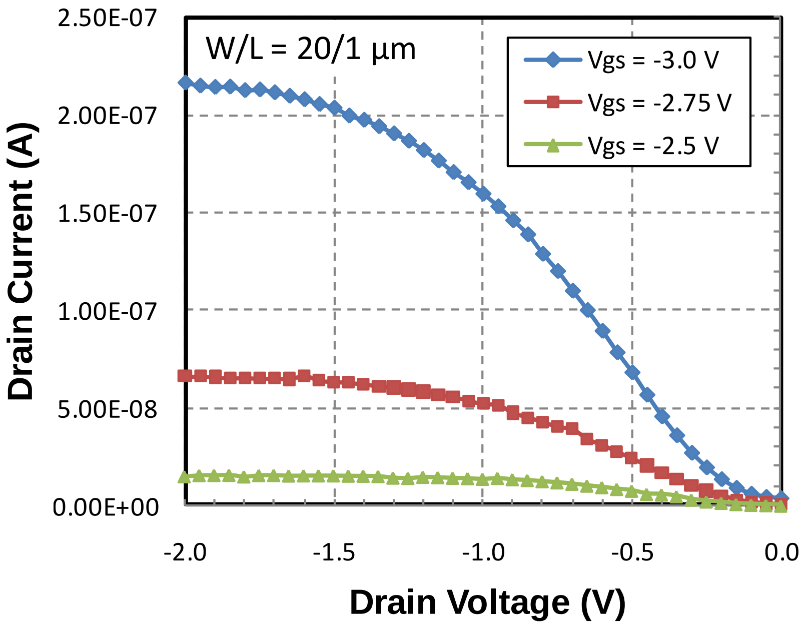


Figure 4.20: Corresponding I_d-V_d measurement of the p-channel control device. Non-linear behavior at low V_{ds} is seen as expected for this device.

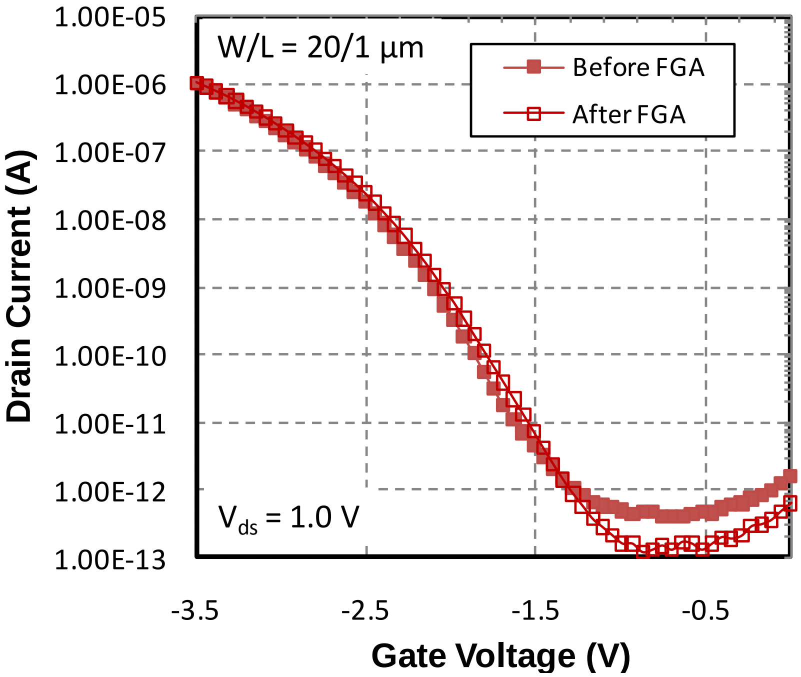


Figure 4.21: Impact of forming gas anneal on the I_d-V_g characteristic of the p-channel control device is minimal.

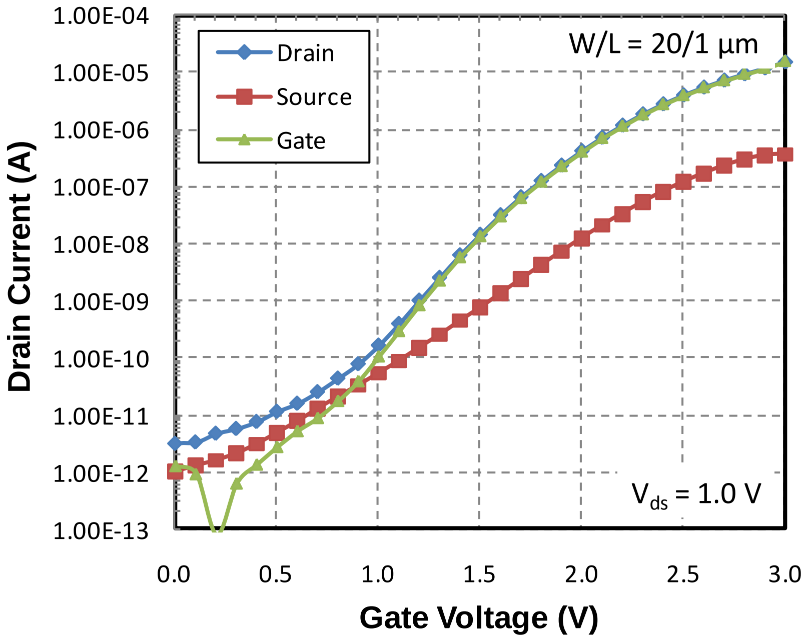


Figure 4.22: Measured I_d-V_g of typical n-channel control device which did not receive pocket implant. In this case, the gate to source leakage is largely overwhelming any potential band-to-band tunneling current in the semiconductor.

Figure 4.23 shows the I_d-V_g representative of the different arsenic pocket dose splits. The results are similar to the gTFETs fabricated in Sematech. The swing is not very steep (approximately 200 mV/dec) and all curves are essentially identical across all dose splits. This suggests that source to pocket tunneling is not occurring in these gTFET wafers as was the case in the Sematech experiment. Increased arsenic dose also does not change the turn on or overlap voltages on these wafers giving further evidence that tunneling is occurring entirely within the source edge region and not in the pocket. The "sudden overlap" steep swing still remains elusive. The failure of both this experiment and those in Sematech to produce steep switching devices suggests that pocket formation via implantation is not the ideal process.

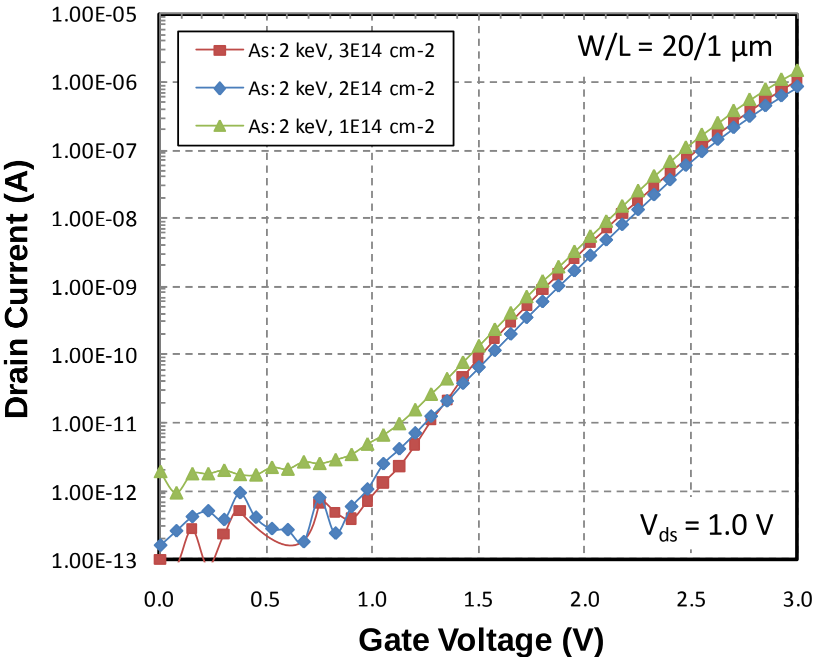


Figure 4.23: Measured gTFET I_d-V_g characteristics. The source implant condition was held fixed at BF_2 10 keV, 5E14 cm^{-2}. These results suggest that source to pocket tunneling is not occurring in these devices.

4.5 Suggestion for Future gTFET Fabrication

Section 4.2.5 discussed two possible reasons why the fabricated gTFETs did not show steep switching "sudden overlap" characteristics. (1) The lateral pocket doping profile from implantation is not abrupt. (2) The poor interface quality between the pocket and gate dielectric does not permit source to pocket tunneling. This results in a gradual "turn on" because of tunneling occurring at the source edge region. Figure 4.24 proposes a process flow that addresses both of these concerns. Doped epitaxial silicon P+ source/N+ pocket/undoped cap layers are grown blanket across the active region. The thin undoped cap (1 nm) serves to improve the gate dielectric interface quality. The layers are anisotropically etched away in half the active region and refilled with undoped silicon. This ensures that the lateral pocket profile is perfectly abrupt.

The remaining process steps are identical to before. Figure 4.25 shows simulation of this structure for various amount of epitaxial refill. This process requires selective doped silicon epitaxial growth, a capability which currently does not exist in the Microlab. Perhaps in a few more months when the newly installed epitaxial reactor is online this structure can be realized by future Microlab researchers. Future collaboration with Sematech or outside epitaxial vendors for fabrication of the proposed gTFET is an option that should be considered as well.

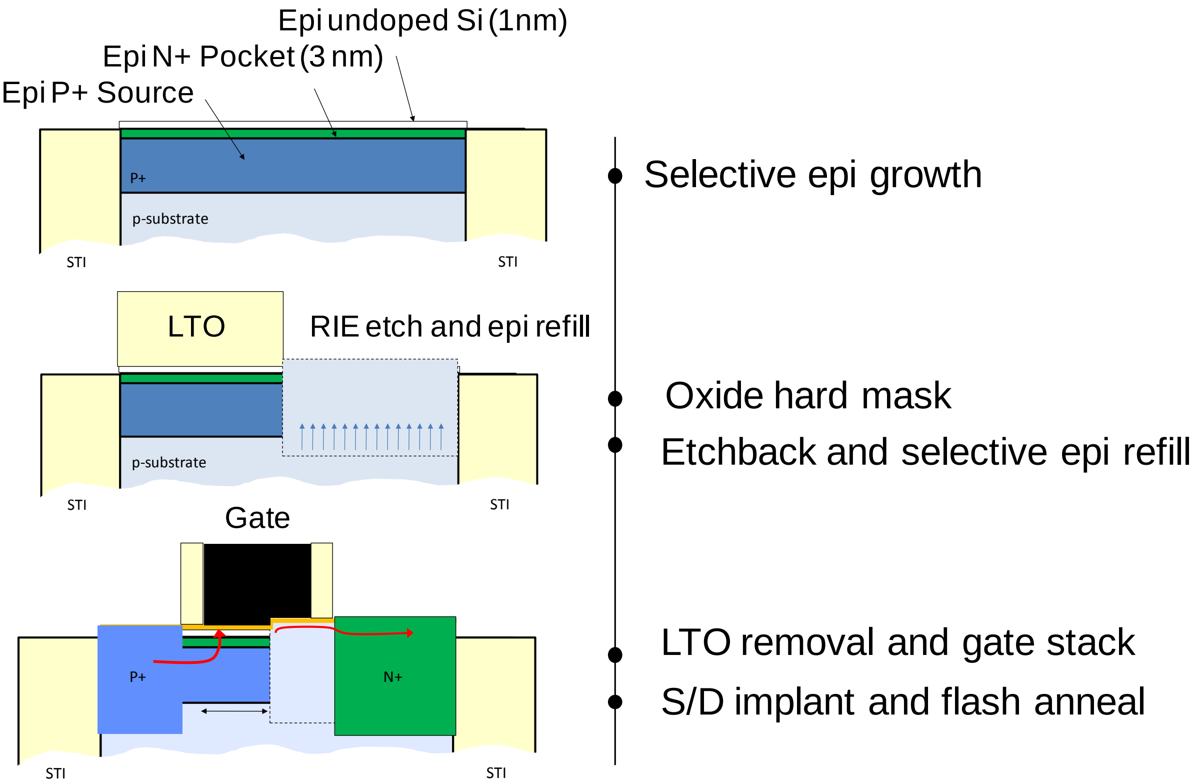
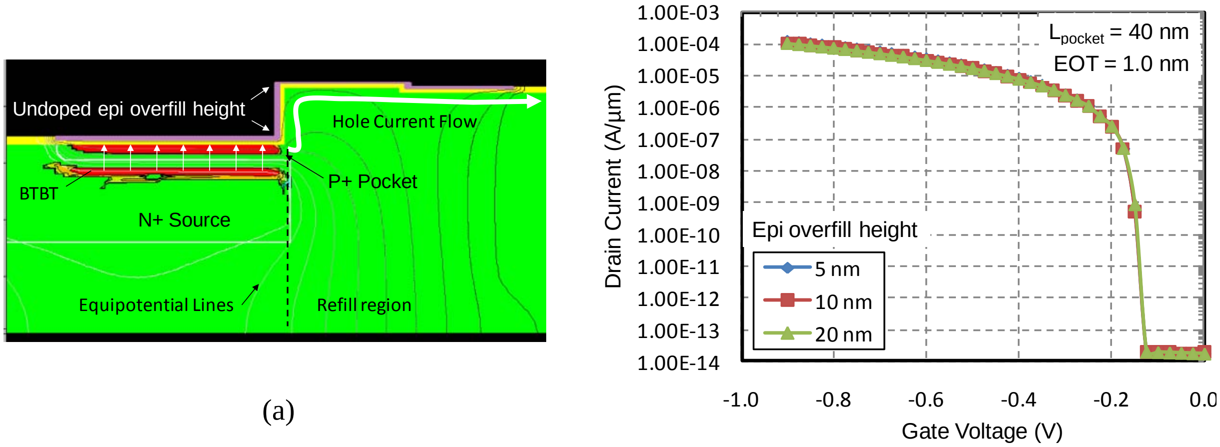


Figure 4.24: Proposed process flow for future gTFET experiment utilizing epitaxial growth. This design addresses dielectric interface quality concerns with an undoped capping layer. The lateral doping profile is made as abrupt as possible from reactive ion etching.



(a)

(b)

Figure 4.25: (a) Simulation output of the proposed structure showing tunneling generation rate and current flow path. (b) gTFET simulation for this structure with varying amount of epitaxial over refill height. The device characteristics are unaffected by variation in this parameter.

4.6 References

[4.1] T. Ito, K. Suguro, M. Tamura, T. Taniguchi, Y. Ushiku, T. Iinuma, T. Itani, M. Yoshioka, T. Owada, Y. Imaoka, H. Murayama, T. Kusuda, "Low-resistance ultrashallow extension formed by optimized flash lamp annealing," Transactions on Semiconductor Manufacturing, vol.16, no.3, pp. 417- 422, Aug. 2003.

[4.2] T. Ito, K. Suguro, T. Itani, K. Nishinohara, K. Matsuo, T. Saito, "Improvement of threshold voltage roll-off by ultra-shallow junction formed by flash lamp annealing," VLSI Technology Digest of Technical Papers, pp. 53- 54, June 2003.

[4.3] T. Ito, T. Linuma, A. Murakoshi, H. Akutsu, K. Suguro, T. Arikado, K. Okumura, M. Yoshioka, T. Owada, “10-15 nm Ultrashallow Junction Formation by Flash-Lamp Annealing,” J. Journal of Applied Physics, vol. 41, pp.2394-2398, 2002.

[4.4] C. S. Kang, H.J. Cho, R. Choi, Y.H. Kim, C. Y. Kang, S. J. Rhee, C. Choi, J.C Lee, "The electrical and material characterization of hafnium oxynitride gate dielectrics with TaN-gate electrode," Transactions on Electron Devices, vol.51, no.2, pp. 220- 227, Feb. 2004.

[4.5] H.N. Alshareef, R. Harris, H.C. Wen, C.S. Park, C. Huffman, K. Choi, H.F. Luan, P. Majhi, B.H. Lee, R. Jammy, D.J. Lichtenwalner, J.S. Jur, A.I. Kingon, "Thermally Stable N-Metal Gate MOSFETs Using La-Incorporated HfSiO Dielectric," VLSI Technology Digest of Technical Papers, pp.7-8, 2006.

[4.6] H. Kato, T. Nango, M. Nakamura, M. Maeda, Y. Ohki, T. Ito, "Effect of post nitriding on electrical properties of high-permittivity hafnium and zirconium silicate films," Properties and Applications of Dielectric Materials Proceedings of the 7th International Conference, vol.2, pp.765- 768, June 2003.

[4.7] S.A. Krishnan, M. Q. Lopez, H. J. Li, P. Kirsch, R. Choi, C. Young, J.J. Peterson, B.H. Lee, G. Bersuker, J.C. Lee, "Impact of Nitrogen on PBTI Characteristics of HfSiON/TiN Gate Stacks," Reliability Physics Symposium Proceedings, pp.325-328, March 2006.

[4.8] J. Huang, P.D. Kirsch, D. Heh, C.Y. Kang, G. Bersuker, M. Hussain, P. Majhi, P. Sivasubramani, D.C. Gilmer, N. Goel, M. A. Quevedo-Lopez, C. Young, C.S. Park, C. Park, P.Y. Hung, J. Price, R. Harris, B.H. Lee, H.H. Tseng, R. Jammy, "Device and reliability improvement of HfSiON+LaO_x/metal gate stacks for 22nm node application," International Electron Devices Meeting, pp.1-4, Dec. 2008.

[4.9] G. Groeseneken, H.E. Maes, N. Beltran, R.F. De Keersmaecker, "A reliable approach to charge-pumping measurements in MOS transistors," Transaction on Electron Devices, vol.31, no.1, pp.42-53, Jan 1984.

[4.10] R.E. Paulsen, M.H. White, "Theory and application of charge pumping for the characterization of Si-SiO_2 interface and near-interface oxide traps," Transaction on Electron Devices, vol.41, no.7, pp.1213-1216, July 1994.

[4.11] M. P. M. Jank, M. Lemberger, L. Frey, H. Ryssel, “Gate Oxide Damage Due to Through the Gate Implantation in MOS Structures with Ultrathin and Standard Oxides,” Conference on Ion Implantation Technology, pp.103-106, Sept. 2000.

[4.12] S. Qin, K. Zhuang, S. Lu, Y. J. Hu, and A. McTeer, “Comparative Study of Self Sputtering Effects of Different Boron-based Low Energy Doping Techniques,” Transactions on Plasma Science, vol.37, no.9, pp.1760-1766, Sept. 2009.

[4.13] S. Qin, Y. J. Hu, A. McTeer, “Advanced Boron-Based Ultra-Low Energy Doping Techniques on Ultra-Shallow Junction Fabrications,” International Workshop on Junction Technology, pp.1-6, May 2010.

4.7 Appendix of Process Details

Pocket First Implantation Process:

	Process Name	Process Specification	Equipment	Comments
1	**Screen oxidation**			
1.1	Piranha clean	Piranha, 120 C, 10 min and 25:1 HF	sink6	
1.2	Screen oxidation	recipe: 1gateoxa 800 C for 3 min	tystar1	Target 3 nm oxide
1.3	Measure oxide	standard ellipsometry	sopra	
2	**Source implant litho**			
2.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
2.2	Exposure	18.0 mJ/cm2	asml	ringFET source implant
2.3	Develop	DUV (Program "1-1-9")	svgdev6	
2.4	Hard bake	program U	uvbake	
2.5	Inspection	SEM	leo	
3	**Pocket and source implantation**			
3.1	Implantation	As 3 keV (dose splits), BF2 10 keV, 5E14 cm-2	core systems	
3.2	Resist ashing	Standard	matrix	
3.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
4	**Dummy LTO gate deposition**			
4.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
4.2	Dummy gate deposition	11sultoa	tystar11	Target 1000A LTO
4.3	Measure LTO	oxide on silicon program	nanoduv	
5	**Dummy gate litho**			
5.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
5.5	Exposure	18.0 mJ/cm2	asml	ringFET gate mask
5.2	Develop	DUV (Program "1-1-9")	svgdev6	
5.3	Hard bake	program U	uvbake	
5.4	Inspection	SEM	leo	
6	**Dummy gate etch**			
6.1	LTO dummy gate etch	Standard MXP-Oxide etch	centura-mxp	Target 900 A removal (leave 100 A behind)
6.2	Resist ashing	Standard	matrix	
6.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
7	**Deep source implant litho**			
7.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
7.2	Exposure	18.0 mJ/cm2	asml	ringFET source implant
7.3	Develop	DUV (Program "1-1-9")	svgdev6	
7.4	Hard bake	program U	uvbake	
7.5	Inspection	SEM	leo	
8	**Deep source implant**			

8.1	Implantation	BF2 10 keV, 2E15 cm-2	core systems	
8.2	Resist ashing	Standard	matrix	
8.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
9	**Deep drain implant litho**			
9.1	Resist coat	DUV 9000 A (Program: "1-2-1")	svgcoat6	
9.2	Exposure	18.0 mJ/cm2	asml	ringFET drain implant
9.3	Develop	DUV (Program: "1-1-9")	svgdev6	
9.4	Hard bake	program U	uvbake	
9.5	Inspection	SEM	leo	
10	**Deep drain implant**			
10.1	Implantation	As 10 keV, 2E15 cm-2	core systems	
10.2	Resist ashing	Standard	matrix	
10.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
11	**Dopant activation**			
11.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
11.2	Anneal	1050 C spike for 2 s	heatpulse4	
11.3	LTO removal	Piranha, 120 C, 10 min + HF dip	sink6	removal all of LTO in HF
12	**Gate stack formation**			
12.1	Piranha clean	Piranha, 120 C, 10 min + HF dip	sink6	remove gate oxide with HF
12.2	High-k dep	standard Al2O3 recipe with 50 cycles	picosun	Target 50 A Al2O3
12.3	Gate deposition	11sdpolya	tystar10	Target 1000 A of poly
13	**Gate litho**			
13.1	Resist coat	DUV 9000 A (Program: "1-2-1")	svgcoat6	
13.2	Exposure	18.0 mJ/cm2	asml	ringFET gate mask
13.3	Develop	DUV (Program: "1-1-9")	svgdev6	
13.4	Hard bake	program U	uvbake	
13.5	Inspection	SEM	leo	
14	**Gate etch**			
14.1	Gate etch	standard 3s OB + ME	lam5	Use endpoint detection to clear etch and end on the Al2O3 gate dielectric
14.2	Inspection	SEM	leo	Inspect gate etch
14.3	Resist ashing	Standard	matrix	
14.4	Piranha clean	Piranha, 120 C, 10 min	sink8	
14.5	Polymer removal	100:1 HF (10s)	sink7	
14.6	Inspection	SEM	leo	Check polymer is removed
15	**LTO deposition**			

15.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
15.2	ILD depostion	11sultoa	tystar11	Target 1500 A
16	**CT litho**			
16.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
16.2	Exposure	22.0 mJ/cm2	asml	ringFET contact mask
16.3	Develop	DUV (Program "1-1-9")	svgdev6	
16.4	Hard bake	program U	uvbake	
16.5	Inspection	SEM	leo	check contact opening
17	**CT etch**			
17.1	Oxide etch	Standard MXP-Oxide etch	centura-mxp	Ensure CT opening are open
17.2	Resist ashing	Standard	matrix	
17.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
18	**Metal deposition**			
18.1	Piranha clean	Piranha, 120 C, 10 min + HF dip	sink6	Use HF to remove native oxide
18.2	Metal deposition	standard 10s sputter etch + Al (2% Si) sputter deposition (standard recipe)	novellus	Target 2000 A of Al
19	**Metal litho**			
19.1	Resist coat	DUV 9000 A (Program: "1-2-1")	svgcoat6	
19.2	Exposure	17.0 mJ/cm2	asml	ringFET metal mask
19.3	Develop	DUV (Program: "1-1-9")	svgdev6	
19.4	Hard bake	program U	uvbake	
20	**Metal etch**			
20.1	Metal etch	MET_Al_mainetch (use endpoint)	centura-met	check visually to see if metal is removed
20.2	Resist ashing	Standard	matrix	
21	**Forming gas anneal**			
21.1	DI water clean	DI water QDR clean and SRD	sink8	
21.2	FGA	Forming Gas Anneal, H2/N2 recipe	tystar18	30 mins

Chapter 5: Ultra Thin Body Green TFET

5.1 Introduction

In Chapter 3 and 4 it was demonstrated that engineering the doping profile of the simple TFET (of Figure 3.1) can result in dramatic improvement in the I_d-V_g characteristics of the device. “Pockets” or thin sheets of charge in this new “green” TFET (gTFET) design allow tunneling to initially occur in a region of large electric field resulting in steep turn on from the sudden overlap of the valance and conduction bands. In this chapter another design of gTFET is detailed that uses ultra thin silicon body on silicon on insulator (SOI) [5.1-5.3] to achieve steep swing by “cutting off” tunneling paths with the buried oxide (BOX). Simulations explore the principle of operation and design space. Fabrication of the silicon UTB gTFET is also detailed.

5.2 Simulations of Ultra Thin Body gTFET (UTB gTFET)

The previous chapters have detailed a novel tunnel transistor design (gTFET) utilizing a heavily doped ultra shallow junction or pocket. When the pocket is engineered correctly, tunneling can occur in a region of high electric field. This results in a very steep turn on characteristic or “sudden overlap effect”, where the tunneling current jumps from zero to a large value when the energy bands initially overlap in the high field region. This is in contrast to the general tunnel transistor or TFET design (from Figure 3.1), where the overlap of energy bands occurs in the low electric field region in the source edge doping gradient resulting in poor turn on characteristics.

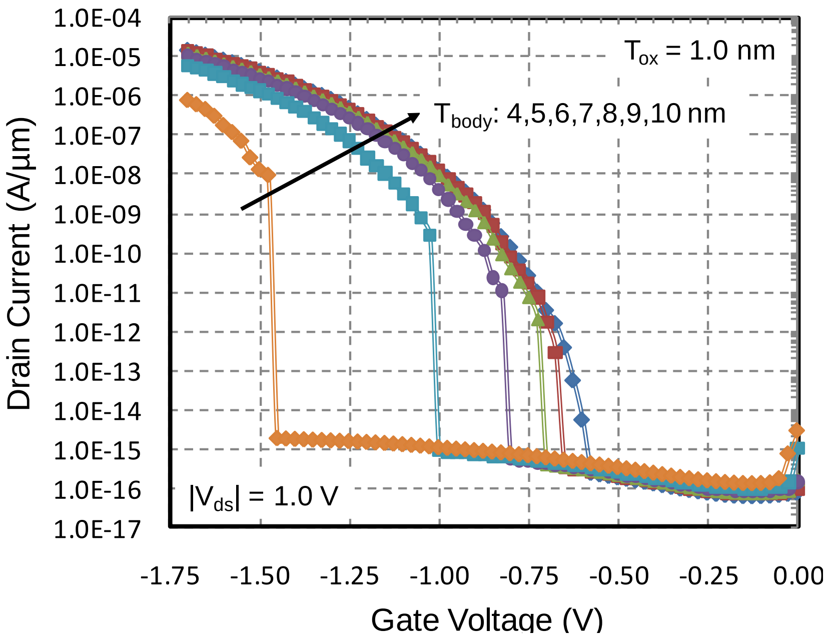


Figure 5.1: Simulation of an ultra thin body TFET where silicon body thickness is varied. Below a certain thickness, steep swing or "sudden overlap" effect is observed.

An interesting effect is shown in the simulation of Figure 5.1. In this case, a generic p-channel TFET (i.e., N+ source and P+ drain) is simulated on silicon-on-insulator with varying silicon body thickness. The simulation tool (MEDICI) and tunneling models/parameters are identical to those used in Chapter 3 (described in Section 3.2.2). For the 10 nm body the typical "gradual" turn on is seen as is expected for the TFET. As the body is thinned further to below 6 nm, the turn on or overlap voltage V_{ov} is increased and a "sudden overlap" steep swing is seen. For the ultra thin body of 4 nm, the swing is comparable to the gTFET of Chapter 3. Correspondingly, this new transistor design is called the ultra thin body gTFET or UTB gTFET. Somehow for ultra thin body on SOI the electrostatics are such to permit "sudden overlap" steep swing.

The principle of operation is shown in Figure 5.2. The electrostatic equipotential contours and band-to-band tunneling generation rate are shown from the simulation of two different body thickness (10 nm and 5 nm) at two different gate voltages (-1.1 and -1.0 V). For the 10 nm body TFET a tunneling path exists at both voltages as indicated by the presence of tunneling generation. Note that as gate voltage is decreased/increased the generation rate contour is pushed away/towards the gate dielectric interface as a result of the electrostatics. However, for the 5 nm device most of the generation is already "cut-off" by the buried oxide (BOX) at -1.1 V. When the gate voltage is decreased by 100 mV pushing the generation rate contours slightly away from the interface, no tunneling path exists and the transistor is completely off in contrast to the 10 nm body device. In this case, the steep swing is caused by the BOX "cutting-off" tunneling generation rate. As the body is made thinner, a larger gate voltage is needed to "pull" the tunneling rate above the BOX explaining the increase in overlap voltage V_{ov}. The electric field is also larger at this new V_{ov} from the smaller silicon body explaining the improved swing.

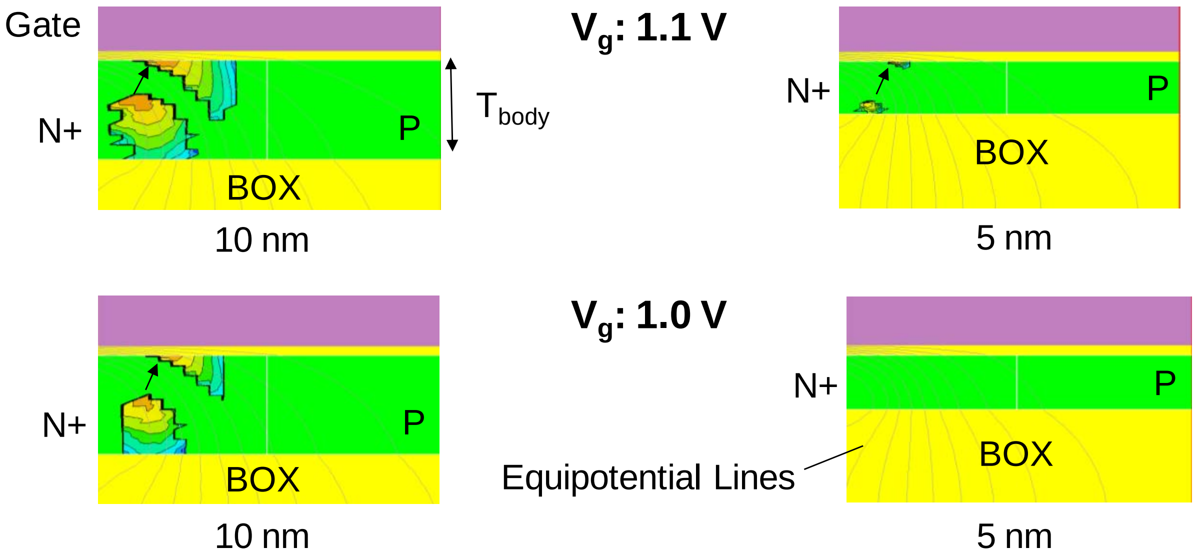


Figure 5.2: Simulation output of ultra thin body gTFET for two different body thickness. The contours represent the tunneling generation rate. For the 5 nm case, when gate voltage is decreased to -1.0 V the generate rate contour is "cutoff" by the BOX resulting in sudden decrease in drain current. This provides another mechanism for achieving steep swing.

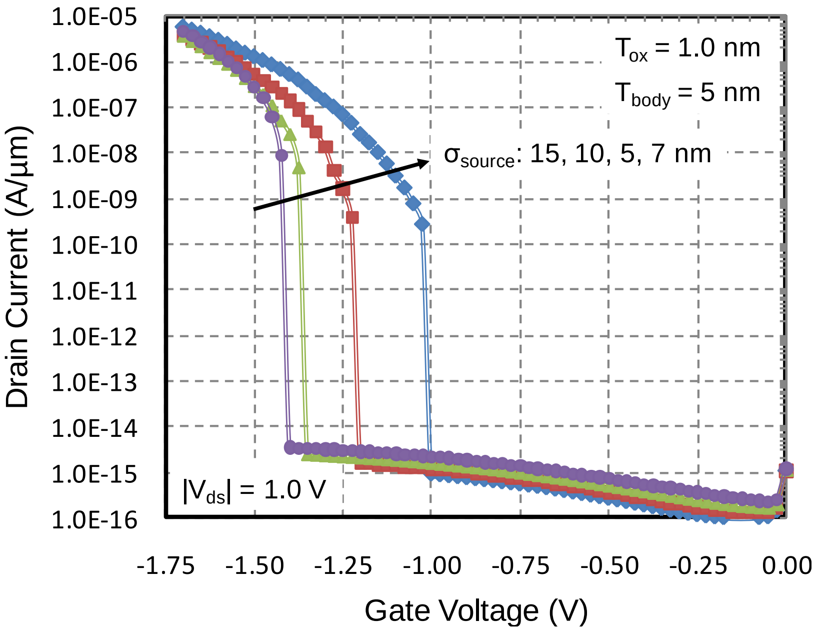


Figure 5.3: Simulation of UTB gTFET for fixed body thickness of 5 nm with varying lateral source profile abruptness. For less abrupt profile the overlap voltage V_{ov} is increased and turn on characteristic improves.

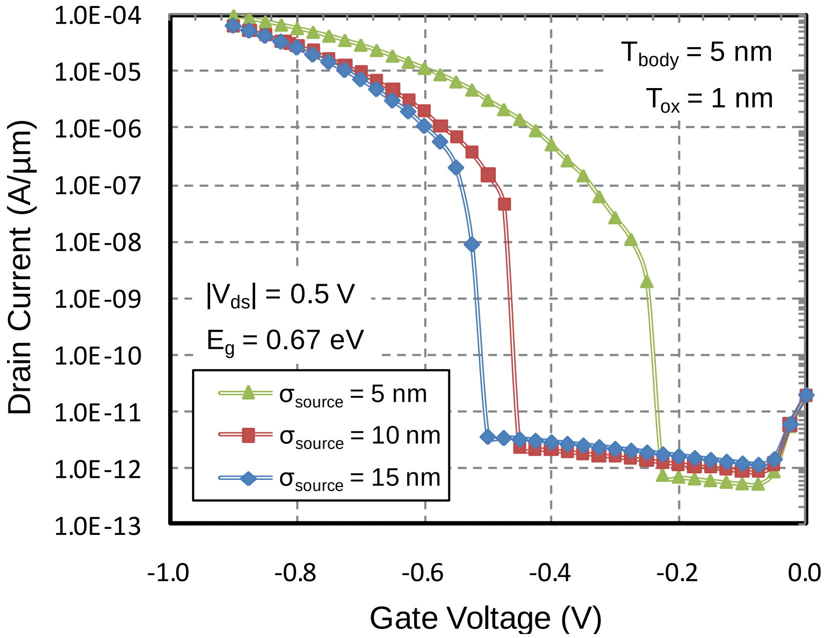


Figure 5.4: Simulation of UTB gTFET on GeOI for 5 nm body. For the least abrupt profile I_{on} of approximately 100 µA/µm is possible for voltage swing of 500 mV.

Compared to the gTFET of Chapter 3, the UTB gTFET has the advantage that fabrication is much simpler since there is no pocket formation involved. However, the UTB gTFET has very large V_{ov} in silicon as seen in Figure 3.1. Significant work function engineering would be needed to lower this value. In addition, the effective tunneling area and therefore drive current is not as large since there is no pocket. Figure 5.3 shows the impact of source doping lateral abruptness on the I_d-V_g characteristic of a 5 nm UTB gTFET. The electrostatics is such that the tunneling direction has more vertical character for less abrupt source. Effectively, the peak generation rate contour is further from the interface, therefore more likely to be “cutoff” by the BOX. More gate voltage is needed to pull the generation rate into the thin silicon body, where electric field is now larger. This provides some explanation on the improvement of the steep swing with less abrupt source profile. Since the V_{ov} is too large for silicon UTB gTFET, germanium or GeOI can be used to both lower V_{ov} and improve drive current. Figure 5.4 shows the results of these simulations.

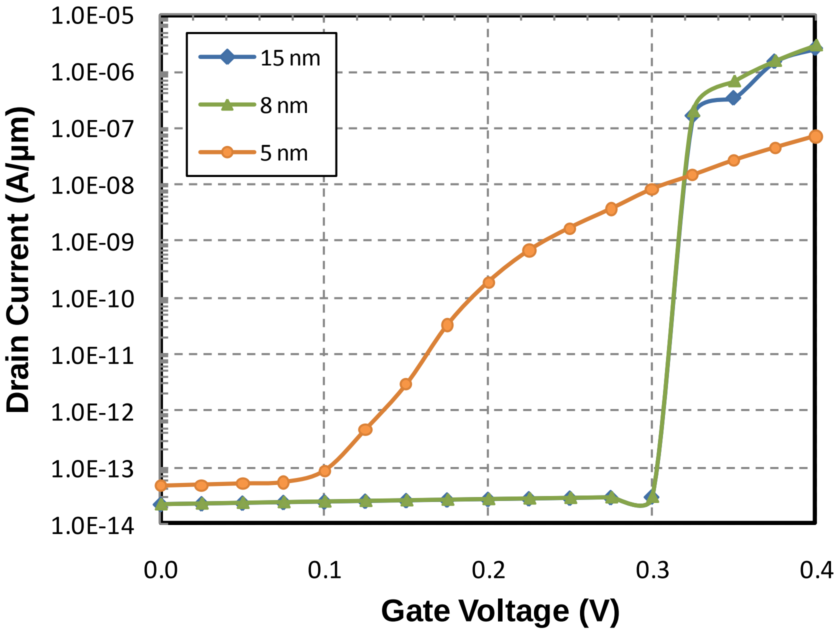


Figure 5.5: Simulation showing the effect of thinning the body thickness of the "pocket" gTFET of Chapter 3 on SOI substrate. When body is too thin, a parasitic tunneling path emerges before source to pocket tunneling is possible.

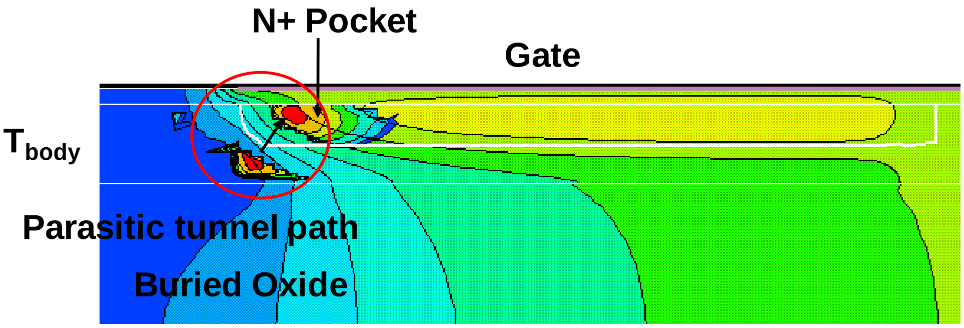


Figure 5.6: Simulation output showing electrostatic potential and tunneling generation rate for "pocket" gTFET on ultra thin body during "turn on".

5.2.1 Combining "Pocket" with the UTB gTFET

Since a pocket of charge is very effective in lowering the overlap voltage in the gTFET of Chapter 3, it is worthwhile to explore its use in the UTB gTFET. Unfortunately, as shown in Figure 5.5, the idea simply does not work. In this case, the gTFET with pocket of Chapter 3 was simulated atop SOI with an ultra thin silicon body. The turn on characteristics are degraded when body is made too thin with the presence of ultra shallow (2 nm) pocket. Figure 5.6 is the simulation output showing a parasitic tunneling path that "turns on" when the BOX is "cutting-off" source to pocket tunneling. This parasitic tunneling path results in a gradual turn on characteristic.

5.3 Fabrication of Silicon UTB gTFET

The fabrication of the silicon UTB gTFET was carried out in Berkeley Microlab. A simplified process flow is shown in Figure 5.7. A small region of silicon near the gate is thinned down before the active layer is defined using a mask as shown in Figure 5.8. The source and drain are formed via ion implantation using drain and source blocking masks.

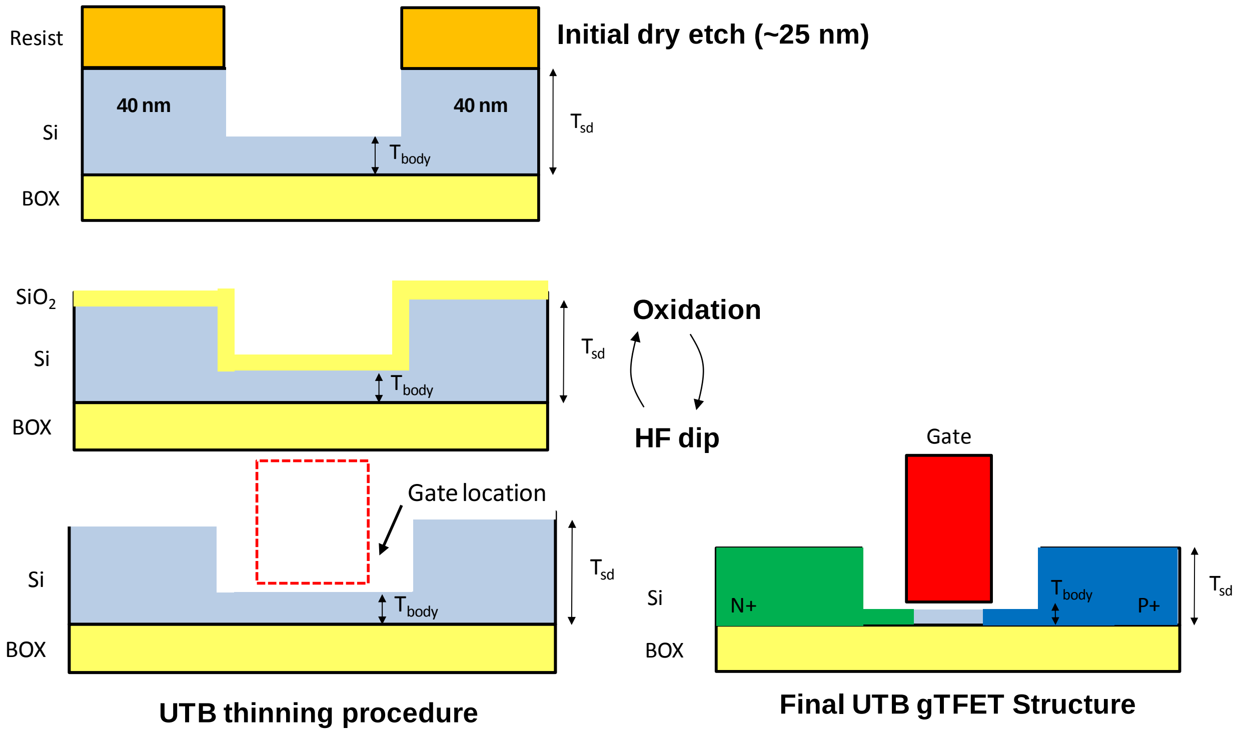


Figure 5.7: Simplified process low for UTB gTFET. The SOI is thinned locally near the gate region prior to gate stack formation. Source and drain are formed via implantation with drain and source blocking masks. The thin body extension region results in additional series resistance, which is likely negligible since drain currents are relatively small for tunneling.

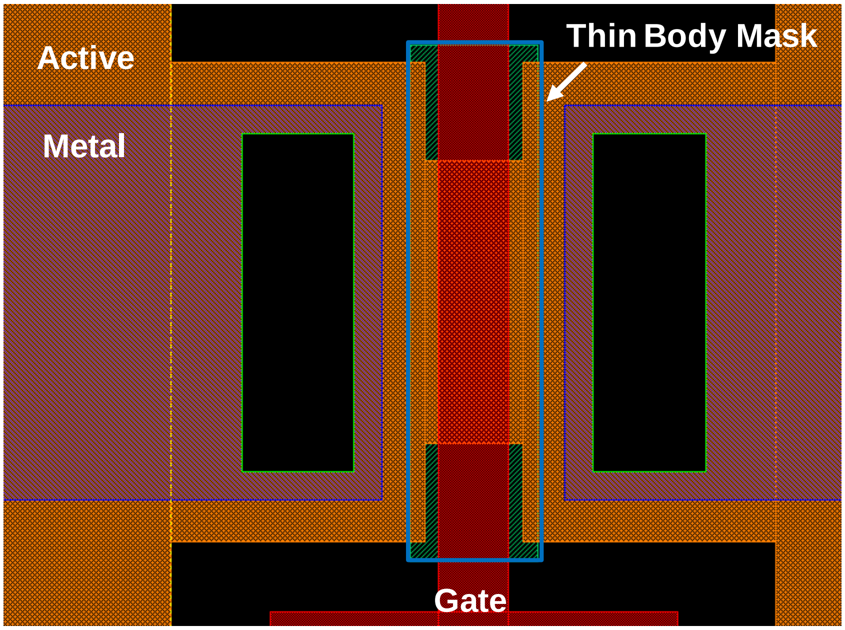
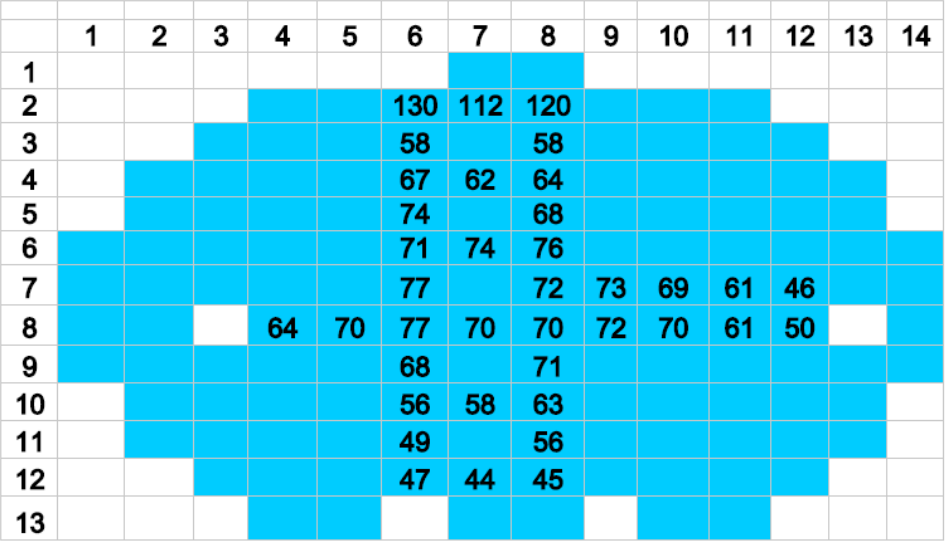


Figure 5.8: Thin body dark field mask which permits the region around gate to be made ultra thin.

5.3.1 Processing Details

A table of the complete details of all processing steps can be found in the appendix of this chapter. The starting materials are 6” SOI wafers from SOITEC with 100 nm silicon body. Atmospheric dry oxidation at 1000° C and subsequent HF dip is used to remove grown silicon dioxide until final body thickness of approximately 40 nm is obtained. Using the body thinning dark field mask (see Figure 5.8), the silicon is reactive ion etched in small opening areas until body thickness of approximately 15 nm is remaining. Afterwards, the wafers are oxidized at 800° C for 3-4 min and then dipped in HF bath to etch away the oxide. This step is cautiously repeated multiple times until desired silicon body thickness is obtained, with each time growing 30-40 Å of oxide. This is to ensure that the entire body is not oxidized away. The thickness can be determined using a thin film measurement tool at various die locations. This is possible because the thinning mask has few test openings that are large enough for die to die thin film measurement. After repeating this cycle numerous times, a body thickness mapping is generated for all wafers as shown in Figure 5.9. The variations across the wafers will result in a variety of body thickness splits during measurement. The die mapping will be used to associate a particular device measurement with its actual body thickness. The locally thinned silicon region is shown under SEM in Figure 5.10. Good surface smoothness is shown, which is important for good quality gate dielectric interface.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1														
2						130	112	120						
3						58		58						
4						67	62	64						
5						74		68						
6						71	74	76						
7						77		72	73	69	61	46		
8				64	70	77	70	70	72	70	61	50		
9						68		71						
10						56	58	63						
11						49		56						
12						47	44	45						
13														

Figure 5.9: Result of wafer mapping after the thinning process. The measurement was done using thin film measurement tool. The units are in angstroms.

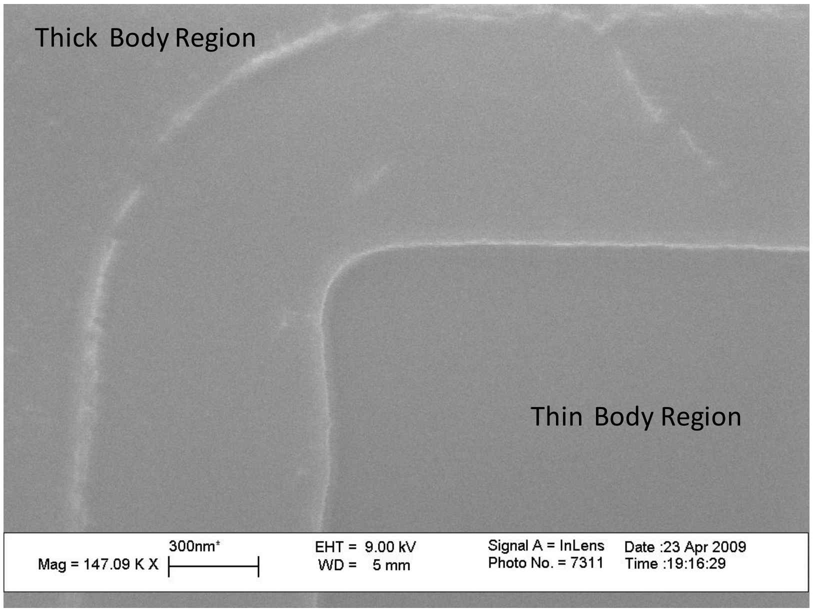


Figure 5.10: SEM showing the surface of the final thinned region looking fairly smooth after successive oxidations. This is important for gate oxide growth.

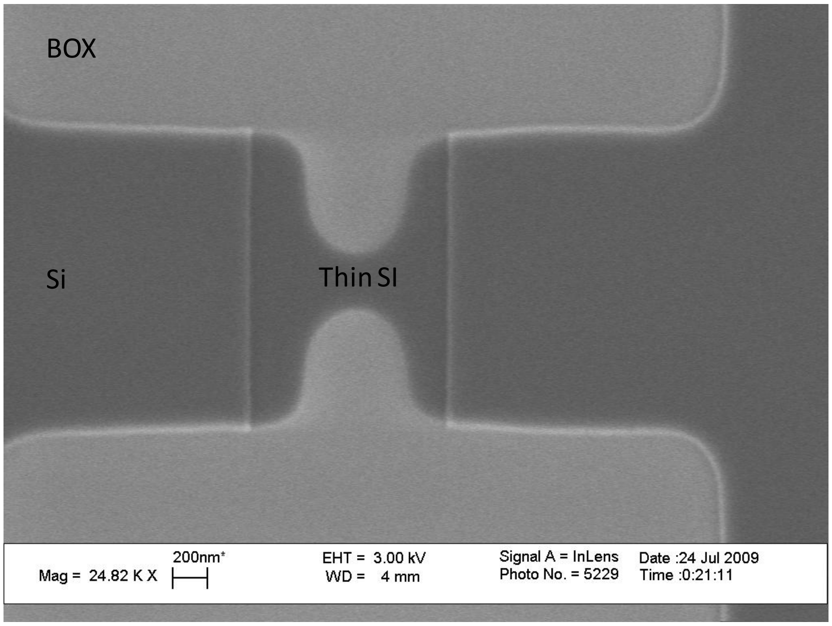


Figure 5.11: SEM of the active region defined after the thinning process. Contrast is seen between the ultra thin body region and the thicker (40 nm) body of the source and drain.

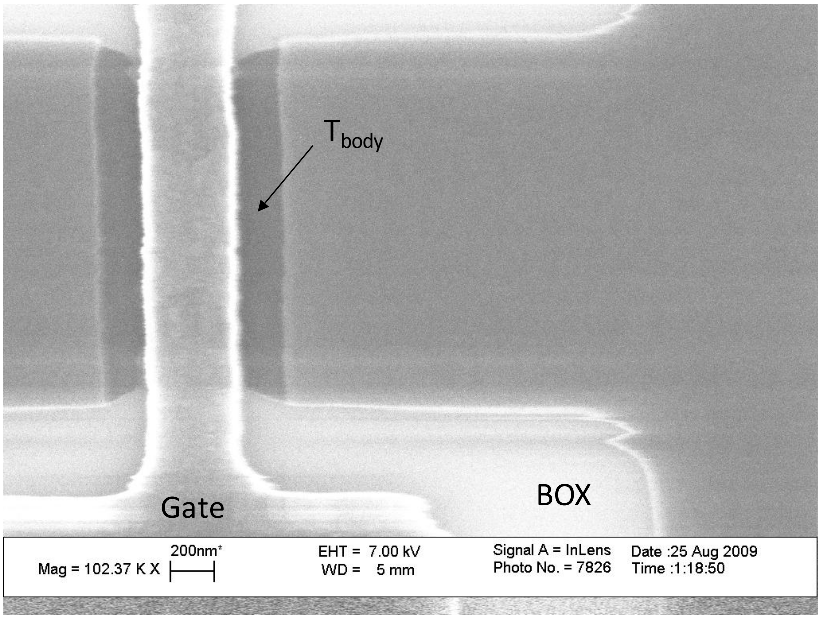


Figure 5.12: SEM after gate etch and clean showing gate landing on the ultra thin body region.

The wafers are then patterned with active mask and then reactive ion etched to define isolated silicon mesas. Figure 5.11 shows an SEM of the active layer with both thick and thin silicon region present. Gate oxidation is then performed at 800° C for 3 min to grow approximately 2.8 nm of silicon dioxide. The wafers are moved immediately into the CVD poly silicon furnace to deposit 1000 Å of in-situ phosphorous doped polysilicon. The gate pattern is applied and polysilicon is reactive ion etched. From experience with prior lots, the over etch step of the polysilicon etch is not as selective as initially thought. The polysilicon gate stringers around the active mesa need to be removed cautiously during the over etch step to avoid removal of the entire source and drain. Figure 5.12 shows an SEM of the gate atop the thin body region in the active region. A 4 nm LTO capping layer is conformally deposited to protect gate edge from ion implant damage using the CVD furnace. The ultra thin body region is also protected from potential sputtering during the implant process. Half active covering implant mask is used to block the drain during BF_2 source implant. The drain side formation is carried out in similar manner using As implant. ILD LTO is deposited and the wafers are rapid thermal annealed at 1000° C for 5 seconds. Contact lithography and reactive ion etch is used to define directly probable contact openings (no metallization required). Standard FGA in a H_2 ambient is then performed to improve the gate oxide interface quality.

5.3.2 Measurement Result

Unfortunately for this lot, all UTB gTFET wafers had source or drain connectivity related problems in that there was no electrical connection between the source and drain. Figure 5.13 shows UTB gTFET I_d-V_g measurement that is representative of all measured devices across all wafers. For this device there is moderate amount of gate-source leakage, but the drain seems disconnected from the other two terminals. An exhaustive search across many different devices, dies, and wafers did not yield a “working” UTB gTFET. The only logical explanation for the formation of “opens” in the device is micro-trenching. [5.4-5.7] This phenomenon occurs during reactive ion etching where ion bombardment can be enhanced near vertical edges (i.e., polysilicon gate edge). This results in increased etch rate near the edge potentially causing localized trenching of underlying material. If this occurred during the gate etch, it is possible the micro-trench could have easily gone through the gate oxide and underlying ultra thin body silicon. In addition, during the initial thin body etch, any micro-trenching could easily have removed the entire body forming an “open”. The risk of micro-trenching was well aware throughout the process. In the past, the polysilicon etch recipe has been tuned by a previous student to remove this problem by reducing the amount of native oxide break through during etching. [5.8] The top down SEM after the gate etch shows no signs of micro-trenching, although the resolution may be the limiting factor in this case.

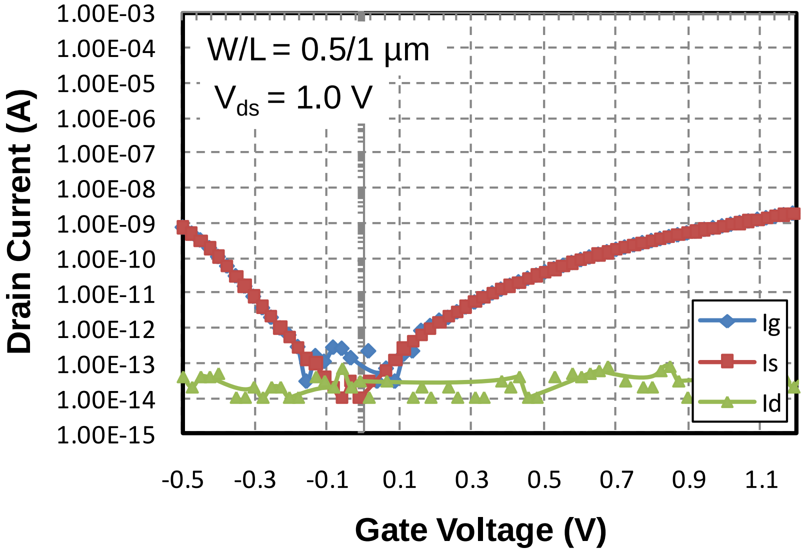


Figure 5.13: I_d-V_g measurement of the fabricated UTB gTFET showing an "open" connection to the drain terminal. In all measured devices either source or drain or both had "open" problems.

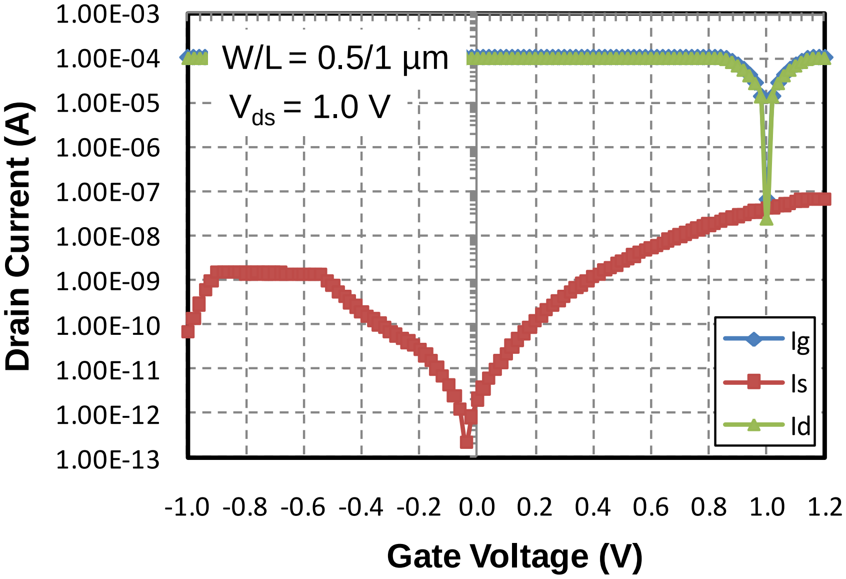


Figure 5.14: Measured I_d-V_g from control device with 40 nm body. A short between gate and drain is seen. The current from the source terminal is likely gate to source insulator leakage. An improper poly stringer removal is the likely suspect for the short.

In addition, the body thinning mask allowed for control thick body (30 nm) TFETs to be made on the same UTB gTFET wafer. As shown in Figure 5.14, these control TFETs suffered from very large gate to drain leakage or a possible short. One possibility is the incomplete removal of the polysilicon gate stringer around the active mesa region, potentially shorting to the drain. This would suggest that the polysilicon over etch step during the gate etch was not enough. However, if a more aggressive over etch was performed it is very possible the source and drain would be etched away entirely as well. Future work on the fabrication of the UTB gTFET needs to focus on these issues.

5.4 References

[5.1] Y. K. Choi, K. Asano, N. Lindert, V. Subramanian, T. J. King, J. Bokor, C. Hu, "Ultra-thin body SOI MOSFET for deep-sub-tenth micron era," International Electron Devices Meeting, pp.919-921, Dec. 1999.

[5.2] Y. C. Yeo, V. Subramanian, J. Kedzierski, X. Peiqi, T. J. King, J. Bokor, C. Hu, "Nanoscale ultra-thin-body silicon-on-insulator P-MOSFET with a SiGe/Si heterostructure channel," Electron Device Letters, vol.21, no.4, pp.161-163, April 2000.

[5.3] P. Xuan, J. Kedzierski, V. Subranmanian, J. Bokor, T. J. King, C. Hu; , "60 nm planarized ultra-thin body solid phase epitaxy MOSFETs," Device Research Conference, pp.67-68, 2000.

[5.4] I. J. Gupta, R. Kraft, S. Krishnan, B. Gale, S. Aur, M. Rodder, T. Laaksonen, "A comprehensive assessment of microtrenching during high density polysilicon etch," International Symposium on Plasma Process-Induced Damage, pp.84-87, June 1998.

[5.5] T. J. Dalton, J. C. Arnold, H. H. Sawin, S. Swan, D. Corliss, “Microtrench Formation in Polysilicon Plasma Etching over Thin Gate Oxide,” Journal of the Electrochemical Society, vol.140, no.8, pp.2395-2401, Aug. 1993.

[5.7] R. J. Hoekstra, M. J. Kushner, V. Sukharev, P. Schoenborn, “Microtrenching resulting from specular reflection during chlorine etching of silicon,” Journal of Vacuum Science and Technology B, vol.16, no.4, pp.2102-2104, Aug. 1998.

[5.8] Y. K. Choi, “Nanofabrication Technologies and Novel Device Structures for Nanoscale CMOS,” Ph.D Dissertation, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, 2001.

5.5 Appendix of Process Details

	Process Name	Process Specification	Equipment	Comments
1	**Thining Litho**			
1.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
1.2	Exposure	19.0 mJ/cm2	asml	Use thining mask
1.3	Develop	DUV (Program "1-1-9")	svgdev6	
1.4	Hard bake	program U	uvbake	
1.5	Inspection	SEM	leo	Check opening sizes
2	**RIE body thinning**			
2.1	Body etch	3s oxide break (OB) through + over etch (OE)	lam5	Target 250 A Si removal
2.2	Measure Si body	poly on oxide program	nanoduv	Check Si thickness
2.3	Resist ashing	standard	matrix	
2.4	Piranha clean	Piranha, 120 C, 10 min	sink8	
2.5	Polymer removal	100:1 HF (15 s)	sink7	
2.6	Inspection	SEM	leo	Check surface roughness
3	**Oxidation body thining**			
3.1	Piranha clean	Piranha, 120 C, 10 min and 25:1 HF	sink6	
3.2	Sacraifial oxidation	recipe: 1gateoxa (various time and temp)	tystar1	
3.3	Oxide removal	25:1 HF until dewet	sink6	
3.4	Measure Si body	Poly on oxide program	nanoduv	Measure into large thin opening. Repeat 3.1 many times till desired thickness. See detailed wafer map.
4	**Active litho**			
4.1	Resist coat	DUV 9000 A (Program:"1-2-1")	svgcoat6	
4.2	Exposure	18.0 mJ/cm2	asml	Active mask
4.3	Develop	DUV (Program: "1-1-9")	svgdev6	
4.4	Hard bake	program U	uvbake	
4.5	Inspection	SEM	leo	Check alignment to thinned region
5	**Active Etch**			
5.1	Silicon etch	3 s OB + OE	lam5	Etch to buried oxide
5.5	Inspection	SEM	leo	Check silicon is cleared
5.2	Resist ashing	standard	matrix	
5.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
5.4	Polymer removal	100:1 HF (10 s)	sink7	
5.5	Inspection	SEM	leo	Check polymer is removed
6	**Gate stack formation**			

6.1	TCA clean	1tca overnight	tystar1	RCA 25:1 bath as well
6.2	Piranha clean	Piranha, 120 C, 10 min and 100:1 HF	sink6	Use 100:1 HF for native ox removal
6.3	Gate oxidation	1thin_ox: 750 C 10 min, 900 C anneal 20 min, 25 min ramp ups, 450 C load/unload temp	tystar1	Immediately load into tystar10 (~2.8 nm growth)
6.4	Gate deposition	10sdplya: 1hr, 1 min, 30 s	tystar10	
7	**Gate litho**			
7.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
7.2	Exposure	18.0 mJ/cm2	asml	Gate mask
7.3	Develop	DUV (Program "1-1-9")	svgdev6	
7.4	Hard bake	program U	uvbake	
7.5	Inspection	SEM	leo	Check gate features
8	**Gate Etch**			
8.1	Gate etch	standard 3s OB + 10 s ME and OE	lam5	Remove fixed amount from OB + ME, remove remaining poly with OE (minimal stringer over etch)
8.2	Inspection	SEM	leo	Inspect gate etch, ensure S/D pad is intact
8.3	Resist ashing	standard	matrix	
8.4	Piranha clean	Piranha, 120 C, 10 min	sink8	
8.5	Polymer removal	100:1 HF (10s)	sink7	
8.6	Inspection	SEM	leo	Check polymer is removed
9	**Oxide cap layer**			
9.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
9.2	Cap depostion	11vdltoa	tystar11	Target 3-4 nm thickness
10	**Source litho**			
10.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
10.2	Exposure	18.0 mJ/cm2	asml	Source implant mask
10.3	Develop	DUV (Program: "1-1-9")	svgdev6	
10.4	Hard bake	program U	uvbake	
10.5	Inspection	SEM	leo	Check alignment to gate
11	**Source implant**			
11.1	Implantation	BF2 10 keV, 1E15 cm-2	core systems	
11.2	Resist ashing	standard	matrix	
11.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
12	**Drain litho**			
12.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
12.2	Exposure	19.0 mJ/cm2	asml	Drain implant mask
12.3	Develop	DUV (Program: "1-1-9")	svgdev6	
12.4	Hard bake	program U	uvbake	
12.5	Inspection	SEM	leo	Check alignment to gate

13	**Drain implant**			
13.1	Implantation	As 10 keV, 1E15 cm-2	core systems	
13.2	Resist ashing	standard	matrix	
13.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
14	**LTO deposition**			
14.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
14.2	ILD depostion	11sultoa	tystar11	Target 1500 A
15	**Anneal**			
15.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
15.2	Anneal	1015 C spike for 5 s	heatpulse4	
16	**CT litho**			
16.1	Resist coat	DUV 9000 A (Program "1-2-1")	svgcoat6	
16.2	Exposure	22.0 mJ/cm2	asml	Contact mask
16.3	Develop	DUV (Program "1-1-9")	svgdev6	
16.4	Hard bake	program U	uvbake	
16.5	Inspection	SEM	leo	check contact opening
17	**CT etch**			
17.1	Oxide etch	Standard MXP-Oxide etch	centura-mxp	Ensure CT opening are open
17.2	Resist ashing	standard	matrix	
17.3	Piranha clean	Piranha, 120 C, 10 min	sink8	
18	**FGA Anneal**			
18.1	Piranha clean	Piranha, 120 C, 10 min	sink6	
18.2	FGA	Forming Gas Anneal, H2/N2 recipe	tystar18	30 mins

Chapter 6: Conclusions

6.1 Summary of Work

This work has focused on researching novel solid state transistors with swing much less than 60 mV/dec. In Chapter 2 the local band-to-band tunneling model was derived from the WKB framework and shown to be identical to the model derived by Kane, which used concepts of time dependant perturbation theory. To derive this closed form expression a constant electric field approximation needed to be made. It was shown that the local model agrees very well the rigorous non-local calculation for tunneling probability if the average electric field across the tunneling path is used. Models for hetero-tunneling were detailed and implemented.

In Chapter 3 a physics based model was developed for a simple tunneling field effect transistor (TFET), which agreed very well with experimentally measured data. This model assumed that tunneling was occurring vertical or normal to the gate dielectric and entirely within the source overlap region. It was shown that the conventional TFET has poor “turn on” characteristic because of the “source edge tunneling” concept. Band-to-band tunneling in the source lateral doping gradient or “effective doping gradient” was treated as various segments each with different source doping contributing to tunneling current independently. The net effect was a “gradual” turn on because tunneling initially occurs in the lighter doped regions near the source edge. An improved design called gTFET was shown to solve this problem by using heavily doped “pockets” or thin sheets of charge to ensure tunneling occurs in a region of large electric field. Simulations have shown that V_{dd} of 200 mV with excellent I_{on}/I_{off} is possible when gTFET is combined with appropriate band gap material.

In Chapter 4 various experimental fabrication attempts of the gTFET were described. An ultra low energy implantation approach was used to define the N+ “pocket” region of the gTFET. Across various pocket dose and energy splits, some promise in swing modulation was shown, however, no devices less than 60 mV/dec were demonstrated. Simulations have shown that the pocket needs to be ultra shallow, heavily doped, and have good gate dielectric interface (i.e., very low D_{it}). These requirements are difficult to meet with low energy implantation approach. A selective doped epitaxial process was proposed to address these concerns.

In Chapter 5 a new mechanism for achieving steep swing was shown in a novel device called ultra thin body gTFET or UTB gTFET. This device uses the buried oxide of SOI to “cut-off” the tunneling path to produce a steep “turn off”. Initial silicon experimental fabrication results were described. Unfortunately, the measurements did not yield working devices resulting from processing challenges.

6.2 Future Directions

It is the firm belief of this researcher that the gTFET can and will be experimentally demonstrated if engineered correctly. The correct process capability needs to be utilized or developed. Therefore, future work needs to primarily focus on fabrication of gTFET and the demonstration of “sudden overlap” steep swing. The process flow suggested in Chapter 4 is shown again below. In this case, pocket lateral profile abruptness can be near perfectly abrupt. Also an undoped silicon cap can address interface D_{it} concerns. This structure also has significant processing challenges, but is a worthwhile direction to explore because the gTFET simulation results are very exciting. It is the hope of this researcher that the device in Figure 6.1 can be fabricated in the near future with demonstration of swing much less than 60 mV/dec.

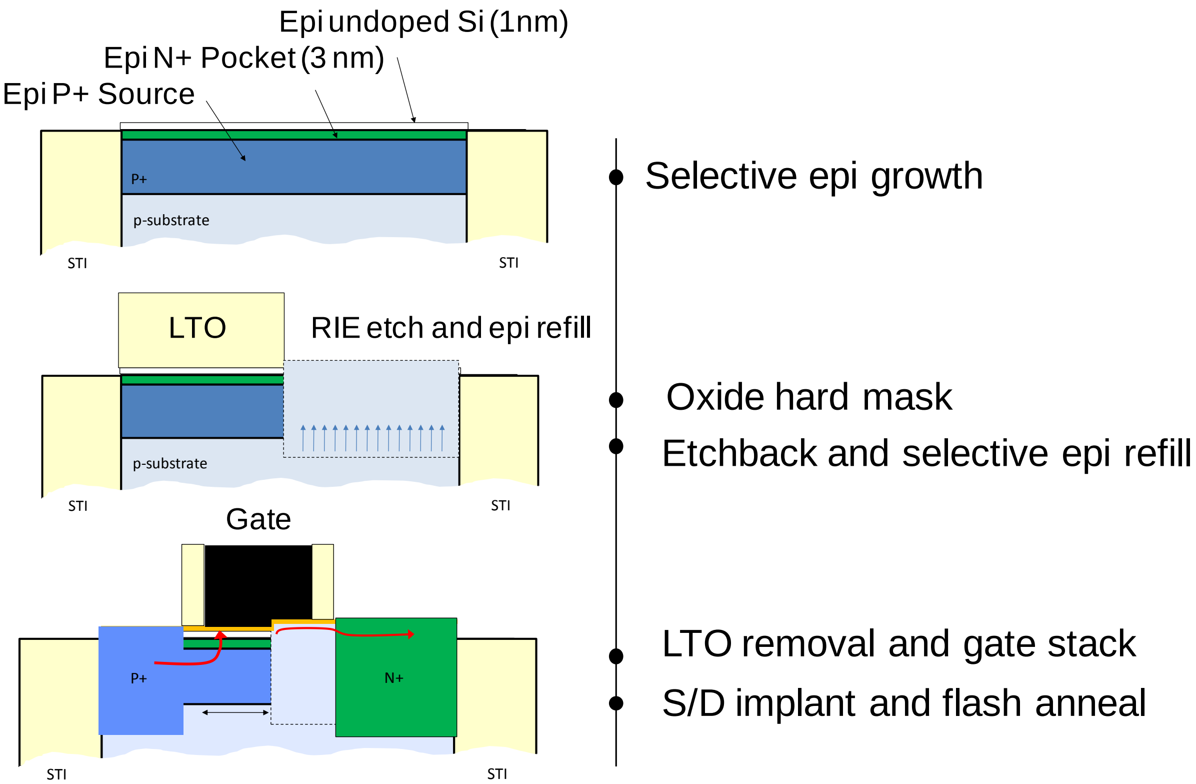


Figure 6.1: Proposed process flow for future gTFET experiment utilizing epitaxial growth. This design addresses dielectric interface quality concerns with an undoped capping layer. The lateral doping profile is made as abrupt as possible from reactive ion etching.